Similarity in Chemical and Protein Space: Finding novel starting points for library design

Zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)
von der Fakultät für Chemie
der Technischen Universität Dortmund
angenommene

Dissertation

von
Diplom-Chemiker
Stefan Wetzel
aus Heidelberg

Dekan: Prof. Dr. Heinz Rehage
1. Gutachter: Prof. Dr. Herbert Waldmann (Technische Universität Dortmund)
2. Gutachter: Prof. Dr. Jean-Louis Reymond (Universität Bern)

Tag der mündlichen Prüfung: 16. September 2009

This work was carried out under supervision of Prof. Dr. Herbert Waldmann at the Faculty of Chemistry of the Technical University of Dortmund and at the Max Planck Institute of Molecular Physiology in Dortmund from January 2005 to June 2009.

dedicated to my family
and friends
for their unconditional support.

“Of course, our failures are a consequence of many factors, but possibly one of the most important is the fact that society operates on the theory that specialization is the key to success, not realizing that specialization precludes comprehensive thinking.”

R. Buckminster Fuller (1895 - 1983)

Table of Contents

General Introduction

Importance of small molecules in modern biomedical research

The discovery and development of small molecule modulators of protein function has a long-standing history in the pharmaceutical industry. Moreover, modern research in chemical biology and systems biology uses small molecule modulators of protein function to study biological systems. Chemical genomics, for example, applies small molecule modulators of protein function to study basic biological processes in cells. Systems biology requires large numbers of specific small molecule inhibitors to study the systemic response to perturbation by modulation of one or more proteins. Therefore, the development of small molecule libraries enriched in biological relevance is a key prerequisite for chemical biology and pharmaceutical research.

Several approaches aiming at the discovery of bioactive small molecules have been invented and are in use today. Whereas the previously very popular screening of natural products, a very rich source of biologically relevant compounds, has been largely abandonend in industry high throughput screening of drug-like libraries, often made by combinatorial chemistry, has taken their place. More than one decade after the advent of high throughput screening, experience shows that these methods yield a developable hit compound in only 30-40% of the cases.[1] To fill the gap, other approaches like high content screening, i.e. the evaluation of cell-based screens according to multiple parameters, and fragment-based drug discovery or yeast genomic screens have been developed. Yet, continuously declining numbers of newly approved drugs tell us that the underlying problem has not been solved by the many paradigm shifts drug discovery experienced over the last decade. This problem is related to biologically relevant chemistry, that is, compounds that are active, suitable for further development and that finally meet the highly complex criteria required for making a marketed drug including efficacy in humans, pharmacology and pharmacodynamics. One of the key questions is, therefore, where to find biologically relevant chemical space in the vastness of the universe of all organic compounds theoretically possible. The field of cheminformatics is actively addressing these questions and provides multiple hypotheses and methods how to best answer them. One possible answer derived from a chemistry background and a pragmatic approach to cheminformatics is presented in this work. This approach has been designed to be intuitively understandable to and useable by educated non-experts, for instance chemists and biologists since understanding often is the first step towards acceptance. A gap exists between experimental and theoretical science whereas in reality both approaches are highly complementary and neither one can exist without the other.

Bridging the gap between chem- and bioinformatics and chemical biology

The computational sciences, especially chem- and bioinformatics, have been developing rapidly due to the availability of high performance computers and arrival of experimental high throughput techniques. These experimental methods, for example high throughput screening or high content screening in systems biology, generate vast amounts of data that can no longer be analyzed manually. The translation of computational results into hypotheses, new experiments and, finally, new science seems often a difficult step, hindered by lack of understanding and communication between scientists on both sides. This work aims at bridging the gap between the scientist at the bench who designs and perfoms experiments and the informatics side of data analysis.

The first part introduces the reader to the concept of chemical space and ways to chart and navigate through chemical space. Promising approaches for the translation of chemical space analysis into chemistry and biology are presented as well. In the light of chemical space analysis and exploration, the development and application of the scaffold tree concept, a hierarchical, structure-based classification of compounds is described. Developed by chemists, the scaffold tree provides a chemically meaningful, intuitive classification of small molecules as opposed to many classification approaches that remain a ‘black box’ to the educated non-expert in cheminformatics. To enable scientists to apply the scaffold tree concept in their daily work, Scaffold Hunter was developed. Scaffold Hunter is a computer program that facilitates visualization of and interactive navigation through large and complex scaffold trees generated from tens of thousands of molecules. Annotation of additional data, for instance biological activity or availability, is easily achieved facilitating quick mining of the results of large biochemical screening campaigns.

Besides the direct analysis of results including the identification of highly active compound classes, the computer can be used to generate new hypotheses. Two approaches are discussed how Scaffold Hunter can facilitate the identification of novel compound classes with desired biological properties. In both cases, the hypotheses generated by the application of Scaffold Hunter were fed back into experimental chemistry and biology and experimental proof for their validity is provided. This exemplifies how truly interdisciplinary research can bridge the gaps between disciplines and integrate computational and experimental science for the mutual benefit of both.

The second part adresses protein structure similarity clustering, a bioinformatics approach for the identification of protein target clusters whose members are potentially modulated by similar compound classes. It describes the evolution from a concept of manual analysis using several available web servers towards a semi-automatic method capable of processing larger data sets. An introduction to the underlying concept and first applications is followed by a detailed description of the method development. The analysis of a larger data set of proteins is

presented together with a critical analysis of its results. Attempted experimental validation of results is given and the implications of its failure are discussed. The developed method can process far more data than the initial manual procedure and relies on well-defined criteria. Nonetheless, it struggles to keep up with the growth of protein structure repositories, e.g. the Protein Data Bank (PDB). Possible improvements of the method are delineated and discussed that may enable the analysis of contemporary data sets with the available computational resources.

In ‘Chemical Biology’ and modern biomedical research, the boundaries between chemistry and biology begin to fade as one seamlessly integrates with the other. The same trend can be observed in the computational sciences, where chem- and bioinformatics methods are fused to address the questions and meet the challenges presented by chemical biology. In light of these developments, the fusion of the scaffold tree and the protein structure similarity clustering is discussed, which may present a viable approach towards merging chemical and biological space.

1 Chemical Space Exploration for Library Design

1.1 Introduction

1.1.1 The concept of chemical space

The chemical space comprises of all organic molecules that are chemically possible[2]; a vast number of molecules. The so-called 'drug space', that is the subspace populated by drug-like molecules, as defined, for example, by the Rule-of-Five[1,[3]], comprises of up to 10^{60} individual molecules.[4]

Similarly to astronomical space that is broken down into galaxies, chemical space can be divided into subspaces defined amongst others by origin or role of chemical compounds. Such subspaces include, for example, drug space populated by the known drugs, medicinal chemistry space comprising all compounds within the Rule-of-Five criteria (sometimes merged with drug space), or natural product space. Natural product chemical space encompasses those parts of chemical space that are populated by natural products. The size of this space is difficult to estimate since it heavily depends on the definition of natural products as well as the sources of information on which the estimate is based. One of the most comprehensive reference works of of natural products, the Dictionary of Natural Products, comprises about 215,000 natural products and analogues in its 17.1 version from 2008.[5] Taking into account J. Bérdy's estimate from 2004 that more than one million natural products were known[6] at that time and the increase of this number within the past years, contemporary natural product chemical space can roughly be estimated to contain several million compounds – a rather tiny but particular fraction of chemical space compared to the 10^{60} molecules in medicinal chemistry space.

The chemical space explored by means of organic synthesis so far is rather small compared to the vastness of chemical space – as of June 2007 about 24 million ring-containing compounds in the Chemical Abstracts' CAS registry database; unarguably the most comprehensive resource of chemical structures published in scientific literature.[7] The diversity of the compounds populating the chemical space explored to date is even lower. In their analysis of about 1,400 marketed drugs from 2007 Siegel and Vieth reported that 15% of these drugs are contained within other drug molecules whereas 30% incorporate other drugs as substructure fragments.[8] The authors use their analysis to obtain drug-based fragment sets for fragment

[1] The Rule-of-Five is an empirical set of parameters to predict compounds which are likely to be orally available drugs. It was derived from known orally available drugs by Christopher Lipinski in 1997. In short, the suitable compounds have a molecular weight below 500, an octanol-water partition coefficient (log P) of less than 5 and contain no more than five hydrogen bond donors and no more than 10 hydrogen bond acceptors. The Rule-of-Five has been widely used in the pharmaceutical industry, i.e., for compound library design, selection of screening compounds and many other purposes.

based drug discovery efforts. These numbers may also reflect the efforts in analogue-based drug discovery (me-too drugs) within the pharmaceutical industry. A comprehensive analysis of the chemical space explored by organic synthesis can be found in the recent analysis of the Chemical Abstracts Services (CAS) registry database by Lipkus *et al*.[7] The authors analyzed the 24.3 million organic molecules within the CAS registry that possess less than 253 non-hydrogen atoms and contain at least one ring. In brief, these compounds break down into 3.4 million scaffolds but only 800,000 frameworks ignoring atom and bond types. Hetero atom containing frameworks were found to comprise mainly of up to 50 non-hydrogen atoms with 50% of the frameworks incorporating 20 to 30 non-hydrogen atoms. The number of ring systems is found to be less than 6 with a sharp maximum at 3. The most frequent 30 scaffolds represent 17.2% (= 4.2 million) of the compounds whereas about 6,500 (0.25%) of the scaffolds are found in 50% of the compounds. Switching from scaffold to frameworks by omitting the atom and bond types, the numbers increase and the 10 and 30 most common frameworks represent 26.1% and 35.7% of the compounds in the database, respectively. Amazingly, 1.3 million scaffolds (50%) occur only once, i.e., relate to only one compound in the repository. Similarily, 47% of all the frameworks describe only 1.6% of the compounds in the database. The reasons for the high similarity found among most of the compounds published to date are subject to further investigations but the general notion that compounds exhibiting high similarity to known compounds might be easier to synthesize[9] may contribute by forming self-enforcing trends towards particular compound classes.

1.1.2 Navigating through chemical space – an overview

The past years witnessed the development of several approaches to charting of and navigation through chemical space addressing different application scenarios. These methods focus either on the molecular properties or on the molecular structure, thereby providing different but complementary perspectives on chemical space. A selection of methods as well as their application to charting of chemical space will be presented in the following paragraphs.

Calculatable molecular properties have been used for decades to chart chemical space. In principle, the position of a given molecule in an n-dimensional property space is defined by an n-dimensional vector comprising of the corresponding molecular property values. It has been shown that molecular property-based methods are able to discriminate between the sub-space occupied by medicinal chemistry compounds, drugs and natural products.[10-12] For visualization, reduction of the dimension of the molecular property vector space to two or three dimensions is necessary, which is achieved by a mathematical transformation named principal component analysis (PCA). PCA generates a new set of basis vectors, i.e. axis of a diagram, comprising of linear combinations of the original n basis vectors favoring those basis vectors that contribute most to the variance of the data set. These linear combinations are called the “loading”. For charting chemical space, the n-dimensions of the molecular properties are then transformed by

PCA to two or three axes. The molecular property vectors of each molecule are transformed accordingly. In the resulting diagram, each dot represents one compound. A scatter plot showing a view on the subspaces of natural products taken from the Dictionary of Natural Products[5], synthetic organic compounds from various compound vendors and drugs taken from Drugbank[13] is shown in Figure 1. This analysis is based on selected molecular properties that are element counts for oxygen, nitrogen and the halides, the number of stereogenic centers, of fused ring systems, as well as the number of aromatic rings divided by the total ring count. Feher and Schmidt found that these properties distinguish between natural products and medicinal chemistry compounds.[10] The resulting diagram shows the subspaces of natural products (see Figure 1a), synthetic compounds (see Figure 1b), and drugs (see Figure 1c). Whereas the sub-spaces of natural products and synthetic compounds overlap to a small extent in this analysis, the drug molecules occupy parts of the natural product as well as of the synthetic chemical space. This finding reflects the close relationships of drugs to the compounds they originated from, e.g. synthetic compounds in screening collections or natural products. Interestingly, synthetic compounds form a small, sharply defined cloud as opposed to drugs and natural products that occupy much larger sub-space. The low discriminatory power of the model may partly result from the reduction to two dimensions only.

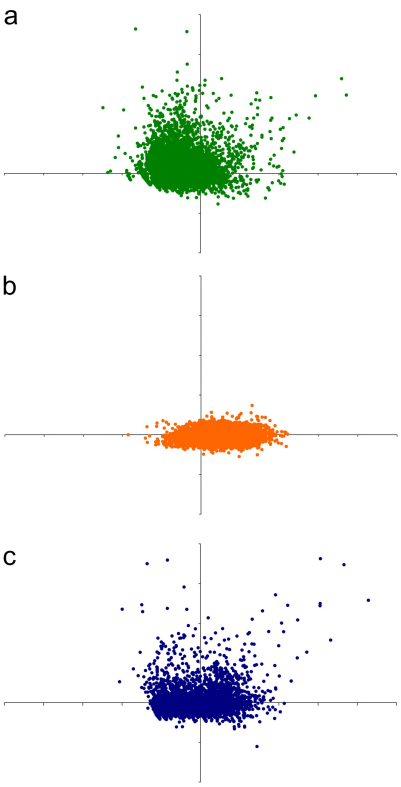


Figure 1: PCA analysis using the molecular properties that were found to differentiate NPs from other compound sets by Feher and Schmidt. Part a) displays 15,000 compounds from the Dictionary of Natural Products, version 17.1 (2008), b) shows about 15,000 synthetic compound picked randomly from various vendor libraries and c) displays a set of 4800 drugs taken from Drugbank. All compounds were classified by the same PCA model and are displayed in similarly sized coordinate systems.

PCA models like the one shown in Figure 1 depend on the data set since their loading, i.e. the axis, promotes those properties that contribute most to the variance of the data set. Thus, any change of the data set that leads to a change in the variance will result in a different model which renders comparisons of different data sets difficult. In 2001, Oprea and Gottfries presented a PCA-based method to charting chemical space that overcomes this problem.[14] Their system named 'ChemGPS' builds a 3-dimensional PCA model based on a set of reference compounds with extreme molecular properties that, hence, occupy the peripheral regions of the chemical space

charted by this model. Compounds are mapped into this reference framework by interpolating their coordinates in relation to those of the reference compound set. This procedure resembles the mode of operation of the NavStar Global Positioning System (GPS) where a network of satellites in geostationary orbits, that is far away from earth's surface, form a reference system from which the position of a receiver on earth can be triangulated using the signal runtime; therefore the name 'ChemGPS'. Oprea and Gottfries selected 423 molecules with extreme molecular properties in medicinal chemistry space that form the reference set of ChemGPS and serve to generate a 3-dimensional PCA model. The loading, i.e. the properties mapped onto each axis, was optimized as well with a focus on chemically interpretable properties, e.g. hydrophobicity, molecular size or molecular flexibility.[15,16]

Larsson *et al.* later applied ChemGPS to chart natural product chemical space.[17] However, a fraction of natural products were mapped outside the reference framework defined by Oprea and Gottfries[14] and, therefore, *extra*polated rather than *inter*polated. This is due to their particular molecular properties that are more diverse than those of medicinal chemistry compounds, which can be seen by the larger extension of the cloud of natural products in chemical space depicted in Figure 1a. To enable charting of and navigation in natural product chemical space, Larsson *et al*. created an optimized reference set comprising of 1779 compounds – four times the number of the medicinal chemistry reference set. Tests of this ChemGPS-NP[18] system with several hundreds of thousands of natural products structures yielded no outliers and generated the 3-dimensional map of natural product chemical space shown in Figure 2.[18-21]

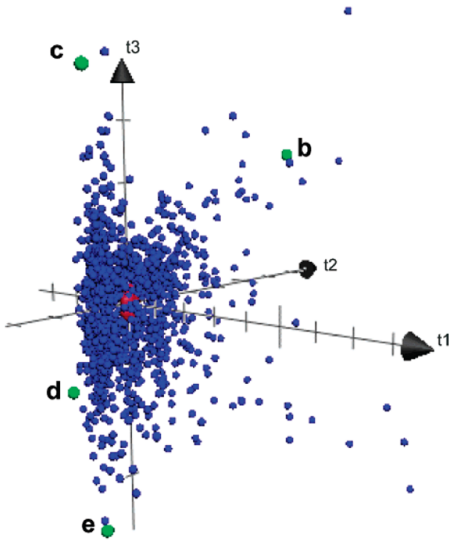


Figure 2: Graphical representation of natural product chemical space obtained with ChemGPS-NP. Each of the blue balls denotes the position of one compound. Reproduced with permission from [18].

In summary, PCA-based approaches are fast and, therefore, well suited for the visualization of the chemical sub-space occupied by large compound collections. The choice of the molecular properties used in the PCA model has a significant influence on the overall analysis since their discriminatory power determines the discriminatory power of the model. However, the application of PCA-based models in the design of chemistry and chemical synthesis is often hampered by the abstract nature of the model complicating the translation of the position of a compound in the PCA model back into chemistry.

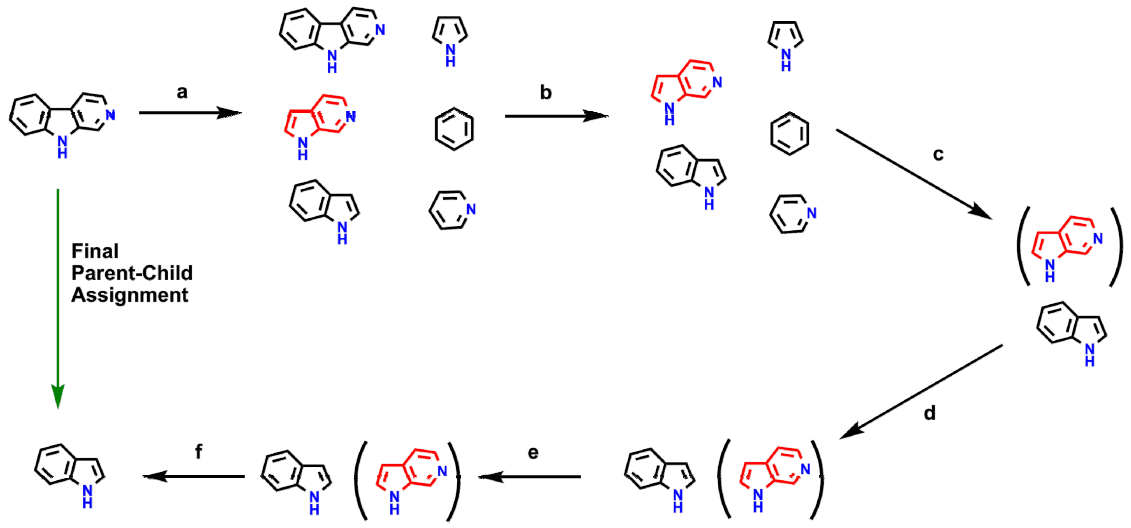


Figure 3: Illustration of the SCONP rule set applied to a sample scaffold. All rules are applied in the sequence shown until only a single parent candidate scaffold remains. Hence, the depicted process ends after step c, the remaining steps are shown for illustration. Scaffolds shown in red are not present in natural products and, hence, not eligible as parent scaffolds. They are only shown for clarity. The steps are: a) identification of all scaffolds whose substructure is present in the initial, i.e., child scaffold. b) Retain scaffolds with less rings than the child scaffold. c) Keep those scaffolds with the maximum number of atoms. d) Retain scaffolds with the minimum number of non-ring bonds. e) Keep scaffolds with the maximum number of hetero atoms. f) Retain scaffolds that occur more often as Murcko scaffold in natural products.

One method aiming at addressing chemical space from the structure side is the “Structural Classification of Natural Products” (SCONP) developed by Waldmann and co-workers. This approach orders the known natural products from the Dictionary of Natural Products by their scaffolds generating a hierarchical classification based on substructure relationships. In brief, the natural product structure is deglycosylated and a chemically meaningful scaffold, comprising of the largest ring assembly and all rings linked to it by one atom, is extracted. This scaffold is then deconstructed iteratively one ring at a time. At each level, exactly one scaffold is assigned as parent scaffold to the larger child scaffold by a set of rules. Only those scaffolds are allowed as parents that are Murcko scaffolds[2,[22]] of compounds themselves. The steps of the algorithm as well as the rules guiding the process are depicted in Figure 3.

These steps are applied iteratively to generate branches of scaffolds as shown in Figure 4. Each molecule is processed individually and forms a scaffold branch. In a last step, all scaffold branches are combined into a tree diagram as shown in Figure 5.

[2] Murcko scaffolds have been defined by Bemis and Murcko and comprise of all rings and linking aliphatic chains of a molecule. They are a widely used representation for compound classes, for instance for classification, definition of structure activity relationships and fragmentation of molecules.

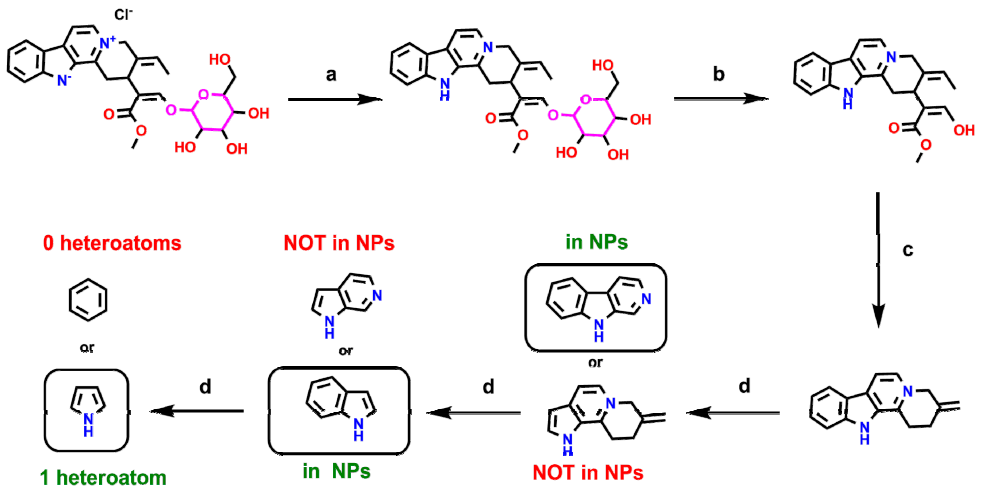


Figure 4: Construction of one branch of the SCONP tree from a natural product structure taken from the DNP. The steps shown are a) removal of charges and counter ions, b) deglycosylation (removal of α-oxy-tetrahydropyran motifs), c) removal of aliphatic linker chains, excluding one atom linkers between rings and ring-based double bonds. d) parent-child assignment according to set of rules depicted in Figure 3.

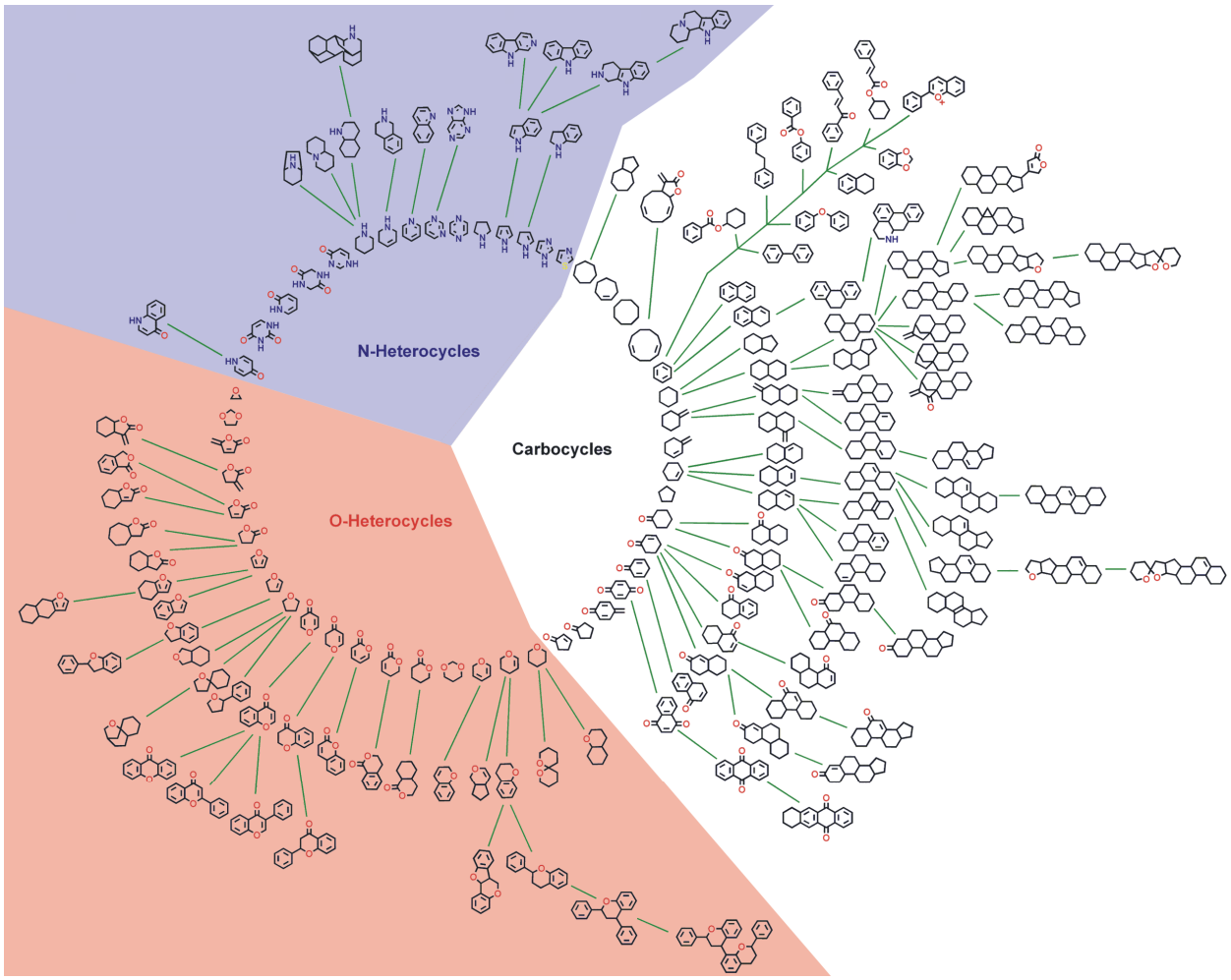


Figure 5: SCONP tree generated from the Dictionary of Natural Products. For clarity, only those scaffolds are shown that represent at least 300 molecules in the data set. Reproduced from PNAS.[23]

1.1.3 Chemical Space Analysis as Tool for the Discovery of New Compound Classes for Medicinal Chemistry Research

Each of the different approaches for charting of and navigation through chemical space provides its individual angle on chemical space. The choice of a particular method strongly depends on the application scenario and the question(s) at hand since all approaches have their strengths and their weaknesses. Additional analyses have been published that utilize the chemical space concept without explicitly charting and visualizing chemical space. This subsection will summarize various ways to explore and exploit the chemical space concept in the quest for novel small molecule modulators of protein function and their corresponding medicinal chemistry programs.

Experience shows that chemical space is mostly devoid of biologically relevant small molecules – just like astronomical space is mostly devoid of matter. Just as stars are stretched out thinly in vast astronomical space, biologically relevant molecules are scattered over chemical space.[2] This explains the typically low hit rates of serendipidy-based screening methods, e.g. high throughput screening, which are typically below 0.1%. The success of a discovery program largely depends on the diversity, quality and biological relevance, that is, the enrichment of biologically active structural motifs, of the compounds examined within the program. [24-28] Naturally, the charting and exploration of biologically relevant part of chemical space as well as the design of compound libraries enriched with structures from these parts are obvious application scenarios for methods charting chemical space.

Biologically relevant structures have been characterized by many criteria empirically derived from retrospective analyses like the rules for drug- and lead-likeness, the Rule-of-Five and others.[29-36] All of these rules are based on the analysis of sets of compounds with proven biological relevance, including drugs, pesticides, herbicides, food additives, medicinal chemistry compounds with proven biochemical or biological activity, and natural products. Natural products are endowed with special properties since they have been selected by nature during evolution for their ability to bind various proteins during biosynthesis, peforming their biological tasks, or bio degradation.[10,37,38] Chemical space analyses can identify the regions of chemical space occupied by these compounds that may also contain other, biologically relevant structures. Without explicitely charting chemical space, Ertl and Schuffenhauer developed the natural product-likeness score, a method to identify natural product like structures which may be enriched with biological relevance.[39] In brief, the authors extracted all two atom fragments and their surrounding atoms occurring in natural products and synthetic chemistry compounds. Then they identified those fragments that were most particular to natural products, i.e. fragments that are over- or underrepresented in natural product structures compared to other compound sets like screening compounds or drugs. The natural product-likeness score of a given molecule can be calculated from the fragments it comprises of. Compounds can be scored in a similar way for

their drug- or lead-likeness. In general, such scoring functions can facilitate descision making related to design, synthesis or acquisition of compound libraries.

An extension of the chemical space concept is the patent space that contains all structures of a given set of patents annotated with further information given in the corresponding patents. Patent space analysis is an important part of drug discovery efforts since the economic value of any development candidate without patent protection is virtually zero.[40] Southall *et al.* analyzed the patent space of 116,550 kinase inhibitors to identify successful chemical strategies for structural modification while retaining their activity.[41] They compared compound series by so-called 'molecular replacements'; a concept developed earlier by Sheridan[42] to analyze drugs for re-ocurring transformations between compound series. In brief, the maximum common substructure, i.e. the largest substructure shared by all molecules, is calculated for two compound series. The remaining substituents are defined as 'R-groups'. The 'chemcial replacement' denotes the exchange of one or more R-groups converting the compound series into one another. Southall *et al.* discovered many chemical replacements that occurred only once in the data set and a limited number of transformations repeatedly used. They were also able to link whole compound series from different companies by sets of chemical replacements, thereby gaining insights into companies' research and development strategies. Common molecular replacements assembled, for instance, in a dictionary may also be used to quickly develop promising me-too drugs from existing compound series.

Waldmann and co-workers developed SCONP to chart natural product chemical space in order to learn about the scaffold structures most abundant in nature.[23] This knowledge could be used in the design of natural product-derived compound collections. Based on the scaffold hierarchy, the authors also proposed a simple, yet effective approach for the structural simplification of natural product scaffold structures while retaining similar biochemical activity. They also presented the application of this concept to a natural ligand of 11β-steroid dehydrogenase type 1 (11β-HSD1) named glycyrrhetinic acid, a compound with a steroid-like scaffold consisting of an assembly of five fused rings. By means of 'brachiation', that is, movement along the branch towards the inner, structurally simpler scaffolds, Waldmann and co-workers identified the octahydronaphthalene scaffold as a promising template for a small compound collection of putative 11β-HSD1 inhibitors. In this case, SCONP provided several smaller scaffolds and the choice was made based on additional information generated from protein structure similarity clustering.[43] Synthesis and biochemical evaluation of a small compound collection yielded several potent 11β-HSD1 inhibitors, indeed. The concept was tested further in the development of phosphatase inhibitors based in the indol-containing branch of the scaffold tree, as well as natural product-derived compound collections targeting other proteins.[44-48]

The methods described so far provide navigation in those parts of chemical space already explored either by nature or by man. Analysis of existing or planned libraries is well possible

and may also provide some guidance for the design and synthesis of compound collections. However, a method extending into the uncharted parts of chemical space, i.e. beyond those parts occupied by known molecules, would be most valuable for the discovery of truly novel small molecule modulators of protein function. In this respect, one may imagine enumeration, that is, automated construction within a computer, of all small molecules chemically feasible. Such attempts have been and are made using either generic aproaches generating molecules from scratch as well as more synthesis-oriented systems using reaction schemes and databases of available building blocks.[49,50] Virtual libraries consisting of billions of compound structures can be generated this way and can later be analyzed by scoring for lead- or drug-like properties. Mapping of these compounds in chemical space together with a reference set of known inhibitors of a protein of interest, for example in ChemGPS[33], can help to identify promising library members. Despite the enourmous advances in computer technology, even today, enumeration of all chemical structures within the medicinal chemistry space, i.e. adhering to the Rule-of-Five, is still impossible.[4,49,51-53] Only smaller subspaces, for instance the chemical space of all small molecules consisting of up to 11 atoms and incorporating carbon, nitrogen, oxygen and fluorine with a molecular weight below 190 dalton have been comprehensively enumerated.[49,52]

The problem of exploration of uncharted regions in chemical space was approached by Van Deursen and Reymond.[54] The authors devised a computer program that ‘morphes’ one molecular structure into another one in a gradual, almost seamless transition. This gradual transition is achieved by repeated cycles of random structural changes and a scoring based selection of structures serving as input for the next round. In brief, several operations including atom type conversion or atom addition are used to generate a set of randomly mutated structures starting from the template. From the resulting structures 10 are chosen based on their similarity to the target molecule. As an element mimicking evolution, another 20 structures are chosen randomly from the remaining molecules and added to the first 10. Together these 30 structures form the basis for the next round of mutation and selection. All chemically possible intermediate structures are kept for later analysis. The intermediate molecules generated by this approach may share properties with both, the start and the target molecule. Reymond and Van Deursen give one example of a transformation starting from AMPA (((*S*)-2-amino-3-(3’-hydroxy-5’-methyl-isoxazol-4’-yl)-propionic acid), an AMPA receptor agonist. As a target, they used a known AMPA receptor antagonist. In principle, the molecular structures resulting from this approach may gradually loose their agonistic activity and acquire antagonistic activity resulting in potential partial agonists and antagonists. The authors did not synthesize nor test any of the proposed structures, so that experimental proof of this hypothesis is still missing.

One could also use this method to morph two molecules addressing different proteins into one another. Some of the resulting intermediate structures may eventually bind to both proteins and, hence, may exhibit a defined poly-pharmacology.

1.2 Aims

The discovery of small molecule modulators of protein function lies at the heart of modern pharmaceutical and bioorganic research. Such molecules are often applied in the study of biological systems using chemical genetics approaches, i.e., modulation of protein function *in vitro* and *in vivo* by small molecules to study the biological effects of this particular protein, e.g. changes in cell phenotype or cell morphology.

Recent years have seen the parallel development of experimental high throughput technologies, e.g. high throughput screening in biochemical and cell-based assays, and of computational methodology, e.g. cheminformatics methods to analyze the resulting vast data sets. Although these fields are highly complementary and would ideally form a symbiosis, a disconnection exists when it comes to the translation of results from cheminformatics analyses and research into new experiments in chemistry and biochemistry and, finally, into new knowledge.

This work is centred at the interface of computational and experimental sciences and aims at the development of approaches and methodology to bridge this gap and close the iterative cycle of experiments, data analysis and the subsequent design of novel experiments based on the data.

The basis for this work was laid by the “structural classification of Natural Products” (SCONP), a substructure-based classification system for charting and navigating chemical space. The aims of this work include:

- The development and maturation of the SCONP approach from a “hypothesis generator” to methodology applicable in biomedical research.
  This evolution should include the extension beyond natural products to other classes of bioactive compounds. In-depth analysis of the set of rules used in SCONP should lead to a new set of rule applicable to many different different applications.

- Development of tools and application scenarios for scaffold-trees in biomedical and chemical biology research.
  This includes the development of an interactive visualization of scaffold trees by means of a computer program to enable experimental scientists, i.e., chemists and biologists without expert training in cheminformatics, to use and apply scaffold trees in their daily research.

- Application of the developed tools in chemical biology research to analyze data sets from high throughput screening for the discovery of scaffolds enriched with biological relevance for the target protein of interest.

  In this subproject, high throughput screening data from biochemical assays should be analyzed in order to identify biologically relevant scaffolds. These scaffolds then form the core of a compound collection that is experimentally tested in biochemical screens to validate the approach.

- Discovery of novel inhibitor structures from natural product-derived libraries using the scaffold trees.

  Novel natural product-based inhibitors should be delineated from analyses of natural product chemical space using the scaffold tree and the visualization tools developed within this work. This includes the identification of a suitable scaffold and potential target proteins, as well as the assembly of a compound collection based on this structural template and biochemical evaluation of these subtances in order to experimentally assess the scope of this application scenario.

1.3 Results

1.3.1 The Structural Classification of Natural Products (SCONP)

Structure-based charting of natural product chemical space is the underlying rationale of the SCONP approach developed by Waldmann and co-workers.[23] In the first step of the scaffold tree generation, a chemically meaningful scaffold comprising the largest fused ring system, all rings linked by one atom to the fused ring system and the ring- and chain-based double bonds, is extracted from the compound structure. In contrast to previous suggestions, ring-based double bonds were added to the scaffold definition and aromatic bicycles were split in such a way that a shared double bond resulted in a double bond at the corresponding position in both child scaffolds (see Figure 6) to retain the scaffold geometry during the deconstruction process. These changes drastically increased the diversity of the scaffolds in the tree diagram as well as its chemical meaning.

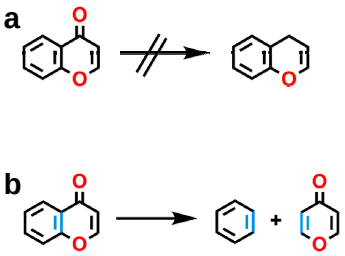


Figure 6: Changes retaining conformation of the scaffolds: a) keep ring-based double bonds as part of the scaffold. b) shared aromatic double bonds create new double bonds at the corresponding position on both resulting scaffolds.

The resulting view on natural product chemical space (see Figure 5) offers a chemically meaningful, structure-based perspective on the most abundant structure types in natural products. The colouring scheme according to heteroatom content in rings reveals a large carbocyclic section and smaller *O*-heterocyclic and *N*-heterocyclic sections. This finding is in line with previous analyses of natural product properties.[10,12,55] The tree diagram also identifies those molecular scaffolds that are most abundant in nature as well as their hierarchical relationships to each other. This knowledge can be directly applied in the design of natural product-derived libraries.

1.3.2 SCONP 2.0: The scaffold tree

One of the most important criteria applied in SCONP allowed only those scaffolds as parent scaffolds that represented chemically meaningful scaffolds of compounds in the data set. The rationale behind this rule was to retain a visualization as close to the natural product space as possible. Not unexpectedly, this method introduced gaps in the scaffold sequences of branches where one or more intermediate scaffolds could not satisfy this criterion. Thus, the composition of the branch for a given scaffold depends on the molecules present in the analyzed data set. This renders comparisons of different data sets, an apparent application of such an approach, almost impossible. This is due to the fact that, most likely, the same scaffold will be assigned to

a different branch for each data set since the branch composition depends on the scaffolds present in the data set.

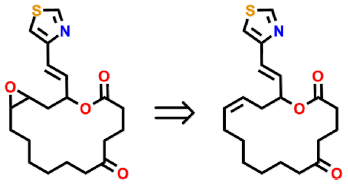
Consequently, together with Ansgar Schuffenhauer and Peter Ertl from Novartis, a new set of rules was devised that no longer required the presence of the scaffolds as molecular scaffolds in the data set.[56] This change effectively removed the gaps and ensured that from any given scaffold, the same branch always results, independent of the data set. Therefore, comparisons of data sets can now easily be made by mapping two scaffold trees onto one another, an application that may be interesting when studying the overlap between different chemical subspaces from a structural perspective.

The rules that guide the scaffold tree generation are more complex than the ones used for SCONP. The new rules were derived from organic and medicinal chemistry knowledge. As in SCONP, the algorithm generates all possible parent scaffolds, i.e., scaffolds with one ring less, and applies the rules in the given order to select or discard scaffolds depending on each rule. Therefore, not only the rules themselves but also the order in which they are arranged has a significant influence on the overall result. All rules are independent of the data set, i.e., the parent assigned to any given scaffold will always be the same as long as the same program is used. The individual rules, their reasoning and mechanism are described in the following paragraphs.

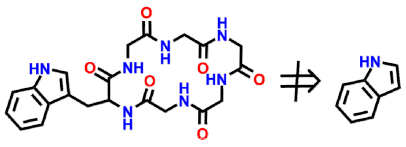
The first rule addresses the removal of heterocycles of size 3, e.g. epoxides. According to organic retrosynthetic procedures, these heterocycles are deconstructed into a double bond between the two carbon atoms of the heterocycle (see Figure 7, Rule 1). This procedure reflects the synthetic approach to epoxides that often proceeds via oxidation of a double bond, e.g. with peracids. If no 3-membered heterocycles are present in the molecule, the rule does not have any effect.

Secondly, macrocycles consisting of 12 atoms or more are preserved since they are regarded as a quite particular motif in many compounds.[57] Many macrocycles, e.g. macrocyclic peptides, adopt well defined three dimensional conformations that facilitate binding to their protein target(s). The three dimensional arrangement as well as the individual molecular interaction are often fine-tuned by the side chains of the macrocycle, i.e., of amino acids, rather than by the backbone. Therefore, the

Rule 1: remove heterocycles of size 3 first



Rule 2: retain macrocycles ≥ 12 atoms



Rule 3: reduce number of acyclic linker bonds

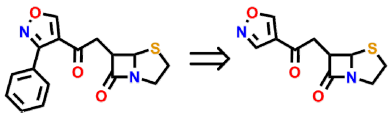


Figure 7: Examples for the rules 1-3 for the scaffold tree generation reproduced from [47].

dissection should retain the macrocycle and start at the sidechains, i.e., with the indole moiety of the tryptophan in the example (see Figure 7, Rule 2). The scaffold tree algorithm does not dissect macrocycles themselves, even if there are no smaller rings left to dissect. Dissection does not pose a good strategy for the classification of macrocycles and an alternative system has been proposed by Wessjohann *et al.*[57] Rule number 3 chooses the parent scaffold with the smallest number of acyclic linker bonds. This ensures a quick reduction of the number of rotatable bonds of the molecules leading to extraction of more rigid core structures. Rigid scaffolds are more likely to possess well-defined and unique interaction patterns. Additionally, aliphatic linkers are natural dissection points in retrosynthesis and often diversified at late stages in compound library synthesis. In the given example, either the phenyl ring connected to the isoxazole or the thiazolidine ring could be removed. Rule 3 ensures that the phenyl ring is removed since this step also removes the single bond to the isoxazole whereas removal of the thiazolidine would remove no acyclic bonds (see Figure 7, Rule 3).

Rule 4 retains unusual structural motifs with characteristic three dimensional conformation, such as non-linear fused or bridged rings and spiro motifs. Such motifs can be defined by two parameters, the number of bonds that are shared between two rings (n_{rrb}) and the number of rings (n_R). The presence and type (linear *vs.* non-linear) of ring fusion, bridged rings or spiro motifs changes both parameters. To determine the degree of presence of such structural motifs, a new parameter Δ is defined that is calculated from n_{rrb} and n_R in such a way, that it is zero for all linear ring fusion patterns (see Table 1, entry 1). This leads to the following definition: $\Delta = n_{rrb} - (n_R - 1)$.

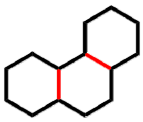
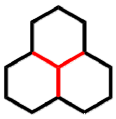
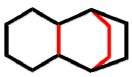
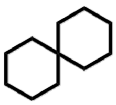
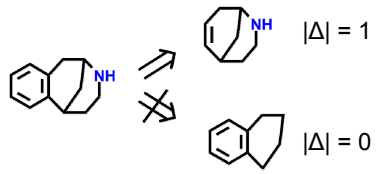
No.	Structure	n_{rrb}	n_R	Δ	|Δ|
1		2	3	2-(3-1) = 0	0
2		3	3	3-(3-1) = 1	1
3		4	3	4-(3-1) = 2	2
4		0	2	0-(2-1) = -1	1

Table 1: determination of parameters for Rule 4 for different structural motifs. The bonds belonging to two rings are coloured in red.

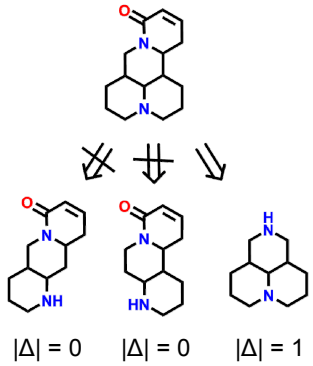
Non-linear ring fusion patterns or bridged ring systems contain more bonds belonging to two rings and, therefore, $\Delta > 0$ (see Table 1, entries 2,3). In the case of spiro systems, the number of bonds shared between two rings is zero yielding $\Delta < 0$ (see Table 1, entry 4). Of course, molecules with rings linked by aliphatic chains would also have a negative Δ. However, rule 3 is applied by first selecting the parent with less acyclic linker bonds and saving the disassembly of the ring fusion patterns and spiro and bridged rings until all acyclic linker bonds have been removed.

Rule 4: retain bridged and spiro rings as well as non-linear ring fusion patterns

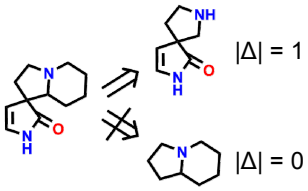
a



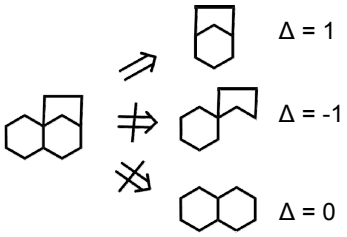
b



c

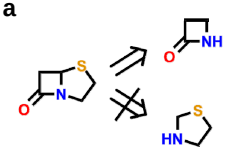


Rule 5: bridged ring systems preferred over spiro systems

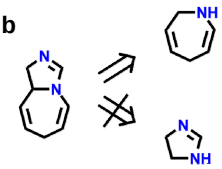


Rule 6: remove rings of size 3, 5 and 6 first

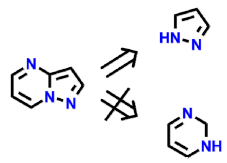
a



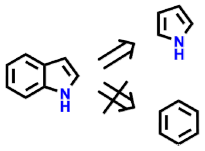
b



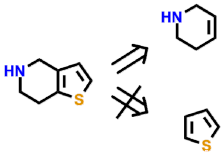
Rule 7: aromaticity must be retained in the parent



Rule 8: remove rings with least heteroatoms first

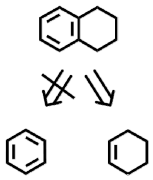


Rule 9: If the number of heteroatoms is equal, keep N > O > S.

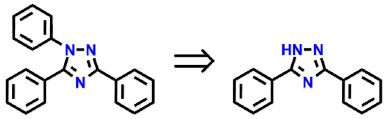


Rule 10: remove smaller rings first

Rule 11: retain non-aromatic rings with priority



Rule 12: Remove rings first where the linker is attached to a heteroatom at either end



Rule 13: if all other rules fail, select the scaffold with the lowest rank in the alphabetical order of the canonical SMILES

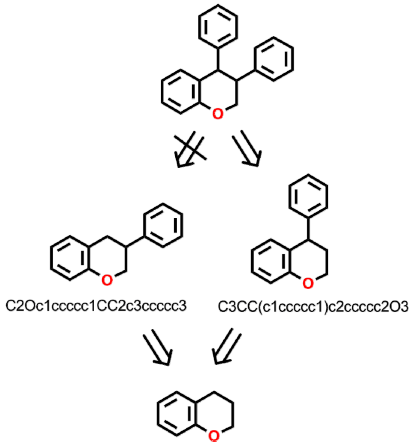


Figure 8: Examples for the rules 4-13 for the scaffold tree generation reproduced from [56].

In the disassembly process, the smaller scaffold with the largest absolute value of Δ ($|\Delta|_{max}$) is assigned as the parent preserving these particular motifs as long as possible. Examples can be found in Figure 8, Rule 4 for the dissection of bridged rings (**a**), non-linear ring fusion patterns (**b**) and spiro systems (**c**).

The dissection of fused systems can eventually lead to the creation of spiro systems as shown in the example in Figure 8, Rule 5. In such cases, it seems chemically more intuitive to retain the ring fusion rather than to create an artificial spiro system that was not present in the corresponding child scaffold. Rule 5 compares the non-absolute values of Δ if two scaffolds show the same absolute value as for the first and second dissection in the given example. To preserve the ring fusion pattern, the solution with $\Delta > 0$ is assigned as the parent scaffold. The third dissection shown in the example is already ruled out by rule four since the absolute value of Δ in this one is smaller than those corresponding to the other two.

Rule 6 removes rings of the size of 3, 5 and 6 atoms first. These rings are found to be abundant in synthesized compound collections and account for most of the commercially available building blocks. Rings of different sizes are far less common and may represent a particular structural motif conserved during series of compounds with similar biological effects. An example is shown in Figure 8, Rule 6 where the dissection of the bicyclic penam scaffold from the antibiotic flucloxacillin is shown. In this case, the beta lactam ring is the decisive moiety present in all antibiotics of this class and is retained due to rule 6.

Rule 7 ensures that the dissection of a fully aromatic system leads again to an aromatic system. As shown in the example (see Figure 8, Rule 7), the dissection produces an aromatic five-membered ring and a non-aromatic six-membered ring. Keeping the latter as the parent scaffold would be chemically counter-intuitive and introduce a non-planar conformation as opposed to the planar conformation of the child scaffold.

The next two rules address the removal of rings containing heteroatoms *vs.* carbocycles. In Rule 8, the ring with the smallest number of heteroatoms is removed first (see Figure 8, Rule 8). This rule reflects the notion that heterocycles may form hydrogen bonds and carbocycles not, which may, therefore, contribute to the binding of the small molecule to its protein target. Should all rings in the scaffold contain the same number of heteroatoms (see Figure 8, Rule 9), a priority of retaining them is established in rule 9, namely nitrogen before oxygen before sulphur. This rule is based on medicinal chemistry experience where nitrogen-containing heterocycles find importance. Sulphur can only undergo very weak hydrogen-bonding and is assigned the lowest priority.

According to rule 10, smaller rings are removed before large rings.

Rule 11 keeps non-aromatic rings with priority over aromatic rings. Thus, in mixed systems (see Figure 8, Rule 11), the aromatic ring is removed first. This rule takes into account the vast abundance of benzene rings in organic molecules and prevents too many molecules being

linked to benzene as parent scaffold. Moreover, non-aromatic carbocycles exhibit a three-dimensional conformation that might be important for the overall molecular shape as opposed to the flat benzene ring.

Rule 12 also reflects synthetic methodology knowledge and removes those rings first, whose linker is attached to a heteroatom at either end. Forming a heteroatom-carbon bond is in many cases preferred over forming a carbon-carbon bond resulting in late stage diversification of compounds *via* attachment of various rings to a scaffold with aliphatic linkers. An example is shown in Figure 8, Rule 12.

In some rare cases, the set of rules may not finally decide which scaffold to assign as the parent scaffold. This may be the case, for example, if a scaffold carries two identical substituents (see Figure 8, Rule 13). In this case, the order of the removal is less critical since both ways converge after the next step, i.e., the removal of the second moiety. Nonetheless, an unambiguous decision has to be made on which ring to remove first. The tie-breaking rule 13 was introduced to solve this problem by assigning the scaffold as parent scaffold that has the lowest rank in the alphabetical order of the canonical SMILES[3,[58,59]] strings. Since there are differences in canonical SMILES generation between different software programs, the outcome of this rule can differ – depending on the cheminformatics software used for canonical SMILES generation.

This rule set has been implemented into “Scaffold Tree Generator”, a publicly available software tool for the generation of scaffold trees. The program is available free of charge from www.scaffoldhunter.com or as supplementary information from the Nature Chemical Biology website: http://www.nature.com/nchembio/journal/vaop/ncurrent/extref/nchembio.187-S4.zip.

1.3.3 Intuitive and interactive navigation through chemical space: Scaffold Hunter

Implications for application of scaffold trees in biomedical research

The scaffold tree algorithm generates a chemically meaningful and intuitive classification of chemical space. However, the result is a data structure that is of little use by itself. Visualization of the classification is absolutely key to enable scientists, i.e., chemists and biologists, to interact with their data and convert data to knowledge and, ultimately, new science - without the need for expert training in programming or complicated software packages. Visualization *via* manually drawn ChemDraw images (see Figure 5) will certainly not be possible for application in every day business. Moreover, the amount of information that can be contained on a sheet of

[3] The Simplified Molecular Input Line Entry System (SMILES) was developed by Dave Weininger in 1988 to encode the structure of a molecule in a text string. However, a given molecule can be represented by many different SMILES strings which led to the development of canonical SMILES where each molecule is represented by exactly one unique SMILES string. One drawback is that canonical SMILES are only unique within the software that generated them but not across different software packages.

paper is limited - by the size and the static nature of the drawing. Dynamic filtering and quick adaptation of visible elements according to certain properties are indispensable elements of large scale data visualization and analysis.

A particularly advantageous solution would be a computer program that automatically generates and displays the scaffold tree for a given set of molecules. Such a program would need to enable users to interact with the data and dynamically update the scaffold tree image according the user's needs; for example with bioactivity values or scaffold selections. One important step would also be data reduction, that is, to visualize only those parts of the data that are needed to address the question(s) at hand. Colour-coding of information, export of relevant results and easy deployment would be key requirements for such a program.

Conceptual design of a suitable computer program

A suitable software tool is user-centred and allows domain experts to analyse hierarchically ordered data from a given database by visual navigation. An intuitive and user-friendly graphical user interface, good documentation, easy deployment, as well as platform independence are required for the successful application in research environments.

The generation of the dataset, in particular the generation of the hierarchical classification and the corresponding data, is kept separate from the visualization software to retain the visualization highly independent of the individual classification approaches, e.g., the scaffold tree. This separation of the classification combined with a generic data structure that can be filled with data generated by various hierarchical classification methods renders the visualization tool as flexible and broadly applicable as possible.

The visualization software comprises graph layout and navigation techniques to explore the chemical space in a structured and task-guided way.

During data analysis, multiple views may be created to visualize data sets resulting from iterated selection and filtering operations. The main view always shows a graph representing the hierarchical relations of the data set whereas detailed views show object properties or selected subsets of the data set. Filtering is an important feature to extract data and create small, focused sets of data allowing visual analysis in a meaningful way. Such visualization for biomedical data had not been developed so far.

The amount of data produced by modern biomedical research requires a modern, perfomant database system. Implementation of the system as a two- or three-tier-model, i.e., a database server and a separate application or a database server, an application server and a client application, respectively, allows easy migration of the system to different database systems accessible via SQL.

Already existing software, i.e., professional database systems or graph-drawing toolkits, is integrated into the project if possible since it will reduce the necessary development work.

However, attention is paid to the source of such modules to guarantee free access and sustainable development over time. Short term solutions seriously reduce the sustainability of the software and re-implementing another such basic module may require a big effort. Wherever possible, open formats and standards should be used to ensure compatibility with standard cheminformatics and structure-drawing software.

Realization of Scaffold Hunter

A computer program as described in the previous paragraph can hardly be developed by one person within a reasonable time.

Therefore, the Scaffold Hunter project was started as collaboration between the Max Planck Institute of Molecular Physiology and Karsten Klein and Prof. Dr. P. Mutzel from the Chair of Algorithm Engineering at the Technical University of Dortmund in spring 2006. It was decided to propose the realization of the first version of the application as the subject of a software project course for advanced computer science students in their 3rd or 4th year of education, which is performed in groups of up to twelve students. The main goal of such a one-year project is to bridge the gap between the introductory examples from undergraduate software courses to real life application development and project management. The students picked their preferred topic from a proposal list; this typically leads to a high level of motivation and commitment to the project goals. In order to guarantee high quality output that is suitable for use in scientific research from a student project, the students committed themselves to strict coding conventions and documentation standards.

The aim of the proposed project group 504 (PG504) was the development of Scaffold Hunter, a program for the visualization of and interactive navigation through scaffold trees. The project requirements and goals were to build working software that retrieves the scaffold tree data from a database system via SQL, visualizes and allows interactive navigation through the scaffold tree, and the export of high resolution data. Since PG504 was part of the educational program, the group should also learn the planning and administration of a software project and provide a full documentation of the program. The project was co-developed and co-supervised together with Karsten Klein from the Chair of Algorithm Engineering. The system and requirement specification was developed in close cooperation with chemists from the Max Planck Institute. Additionally, input from chemists and cheminformaticians with a pharmaceutical industry background was gathered in order to ensure close contact to real-life problems and future users already in the early design stages of the project. Early prototype versions of the system were then given to domain experts who served as test users and gave valuable feedback that helped to improve the software in multiple iterations.

Scaffold Hunter was largely coded within PG504 by its participants, namely Arbia Ben Ahmed, Anke Arndt, Philipp Büdenbender, Vanessa Bembenek, Adalbert Gorecki, Nils Kriege, Sergej

Rakov, Michael Rex, Gebhard Schrader, Henning Wagner, André Wiesniewski and Cengizhan Yücel.

ScaffoldTreeGenerator: open source implementation of the scaffold tree algorithm

ScaffoldTreeGenerator is a program that builds the scaffold tree database according to the set of rules published and described in section 1.3.2. Scaffold Tree Generator has mainly been programmed by Steffen Renner and was later modified within the work for this thesis, in particular errors were removed and the database module was adapted to Scaffold Hunter. Nils Kriege, a computer science student from PG504, added a graphical user interface that increased usability for non-expert users. The Scaffold Tree Generator, as the program is called, uses the Chemistry Development Kit (CDK)[60,61], a cheminformatics toolkit for Java that offers many cheminformatics methods for manipulation and analysis of chemical structures. It automatically connects to a MySQL database system and creates and fills a scaffold tree database from an SD file, a standard file format for molecules whose definition is publicly available[62] and that is supported by most standard chemistry software including the drawing programs Isis Draw[63] and ChemDraw[64] and CDK. Scaffold Tree Generator automatically builds the scaffold tree hierarchy and stores it in an SQL database. It can also calculate numerical properties for each scaffold that are averaged over the corresponding property values of the molecules represented by each scaffold. Scaffold Tree Generator also allows incremental extension of existing database, i.e., the mere addition of new molecules to existing trees.

Scaffold Hunter software design and technical requirements

The design of Scaffold Hunter included the following steps: database design, database interface design, functional design and visualization and layout. The design was kept very modular to facilitate maintenance and extension of the program.

The Scaffold Hunter database was designed in the first step. To facilitate large scale storage and fast retrieval of data, a professional database system like Oracle, DB2 (IBM) or MySQL is used. Since Scaffold Hunter's anticipated user community spans the whole range of biomedical research environments from academic groups *via* small and medium sized companies to big pharma, MySQL was the database system of choice. Many academic groups and institutes are running MySQL instances, it is relatively easy to install, offers a wide range of performance from the installation on a local PC to dedicated database server farms with thousands of processors. The costs of MySQL are quite low – it is free for academics and affordable for industry.

The MySQL database is connected to Scaffold Hunter using standard SQL commands, a query language shared by almost all professional database systems. SQL in combination with a dedicated module for database access minimizes the effort needed to adapt Scaffold Hunter to

another SQL based database system, e.g. Oracle, thereby increasing portability to different software environments.

The design of databases usually adheres to the concept of “normalization”, i.e., the storage of every bit of information at only one specific position from where it is always cross-referenced. This principle was implemented in the design of the Scaffold Hunter database as well, as can be seen in Figure 9. The definitions of the different numerical and string properties, for example, are stored in a table separate from the values for these properties. In the design of the database, the idea of creating a flexible visualization tool for different kinds of hierarchically classified data irrespective of the classification algorithm was strictly pursued (see Figure 9a).

The individual entities forming the basic data set that is classified, i.e., the molecule data (orange), are stored in three tables holding the basic data (structure_data), the numerical properties (structure_num_properties) and the string properties (structure_string_properties). The generic storage of properties as numerical values and text leads to the ability to store virtually all kinds of annotated data. These data are clearly separated from the classification data, i.e., the scaffold data (green). In a similar scheme, basic data (scaffold_data), numerical

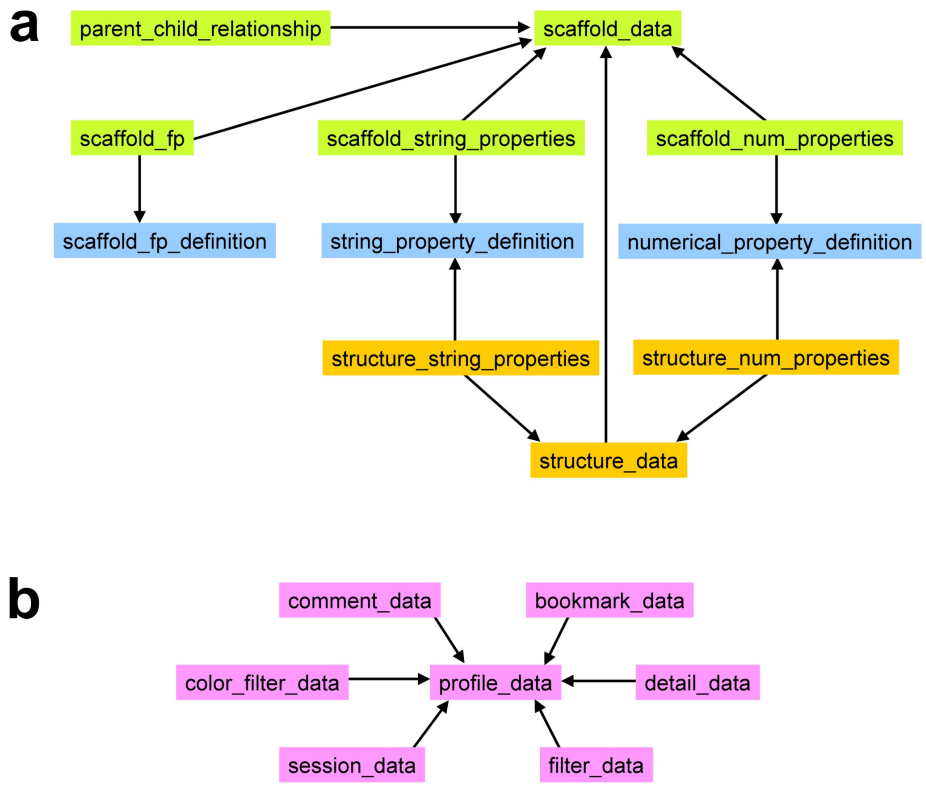


Figure 9: Database diagrams for Scaffold Hunter a) The scaffold tree data is store in databases relating to scaffolds (green), structures (orange) and definitions (blue). The arrows indicate cross-referencing in the direction of the cross-referenced table b) User profiles are also stored in a separate database. All the stored user settings refer to the user profile (profile_data) as central table.

properties (scaffold_num_properties) and string properties (scaffold_string_properties) are stored separately. Definitions of properties are also stored in separate tables since they describe the properties stored for both, molecules and scaffolds. All these data are cross-referenced as shown in the diagram by arrows. The arrows point in the direction of the table that is cross-referenced.

User-related data, for example, selected filters, bookmarked scaffolds etc. can also be stored in a separate user profile database (see Figure 9b). This ensures access to the personal user profile at every computer in environments where computers are shared between people. In this database, everything relates to the user profile that stores all the basic user account information. The user profile can also be stored in an Extensible Markup Language (XML) file in the application data folder on the local hard drive instead of the central database.

Scaffold Hunter users are able to share their work with others by simply exchanging a session file containing all relevant application settings. This way, the visualization obtained by one user can be reproduced by another user, simply by importing the session file and applying it to the same database. The same mechanism can also be used by a single user to restart his work at the same point where the work was interrupted. Scaffold Hunter also includes a number of export capabilities, for instance, high resolution image export, that generate personalized visualization output for presentations and publications.

A particular challenge presents the fact that the software will be run on existing standard hardware and multiple platforms, e.g, Windows, MacOS and Linux. To meet this challenge, it was decided to implement the software using the Java Webstart technology. This technology enables the application to be deployed from a web server to a wide variety of platforms, but can also be used for a purely local installation where the program will be intermediately stored on the client computer. Besides a Java Runtime environment and the database system, no special hardware or software is needed. The Java Runtime environment is standard software already installed on most computers.

The application itself consists of three main modules. The first one is the data access module that is responsible for the database interface and retrieval of data, including filtering mechanisms. The second module, the data visualization module, comprises the canvas visualization, layout algorithms and the navigation techniques. ‘Canvas’ in this respect describes the area on a screen where images are drawn by the program. The third module is the user interface module, which is responsible for user interaction, export capabilities and the view coordination. This design was chosen to create clearly separated modules as described above and to allow clear and simple interfaces. It follows the “Model-View-Controller” pattern which is a design paradigm that breaks an application into three parts to separate data handling and user interface. The “model” part represents the data and rules from the application domain and manages the state of the software, the “view” part allows visualization of model data, and

the “controller” module is responsible for processing events and user actions and triggers changes and updates on the model.

The interaction with large hierarchies is a well-studied problem in the field of information visualization. Therefore state-of-the-art information visualization approaches are implemented in the tool and a combination of intelligent filtering, highlighting, and navigation and layout algorithms from automated graph drawing will be used to meet the requirements laid out before. To keep the memory requirements from the graphical depictions of the scaffolds reasonable, the respective scalable vector graphics (SVGs) are loaded into memory on demand when they need to be displayed on screen and cached for further use with a “least recently used” strategy.

Third party software

Several third party software components were used that are freely available:

Piccolo Toolkit: The visualization builds on the piccolo toolkit which provides a zooming graphics interface, a canvas with object management and event handling, a scene graph model to manipulate objects on the canvas, and also animation features. Even though Piccolo also provides automatic layout features, it was decided to implement additional graph layouts that are specifically suited for visualizing the hierarchical information from the scaffold tree.

MySQL: The database interface is currently implemented using the MySQL database system. The implementation could be easily changed to change the database software to connect to an Oracle database via JDBC.

Batik SVG Toolkit: The SVG rendering and export is done using Batik, a Java-based toolkit that provides modules to generate and manipulate SVG images.

1.3.4 Scaffold Hunter: interactive visualization of chemical space

This sub-section will explain the layouts implemented in Scaffold Hunter as well as its user functionality.

Technical Implementation

Scaffold Hunter was completely developed in Java to ensure a maximum of platform independence. The program has been shown to work under Microsoft Windows, Apple MacOS and different Linux systems. To maximise the sustainability of the Scaffold Hunter project, a concern with all academic software development projects,[65] a dual strategy for extension and maintenance of the software was pursued. On the one hand, development will be continued by the intial developers but on the other the user community will be integrated into these efforts. Hence, Scaffold Hunter was made available under an open source license with its source code. It is available free of charge via its website http://www.scaffoldhunter.com.

As discussed in the section above, the Piccolo toolkit has been used for the basic functions of graph drawing. Chemical structures are stored as SVG images due to their small size and scalability, i.e., a SVG image will have smooth edges at any chosen zoom level. The rendering of SVGs to bitmap images for display is performed using the Batik toolkit.

Graph Layout

The graph defined by the scaffold hierarchy needs to be laid out in an appealing way that at the same time facilitates the exploration process. Depending on the goal of the exploration process, different layouts may facilitate different tasks. Automated graph drawing deals with the layout of relational data arising from many different application areas. Graph drawing provides sophisticated algorithms to lay out graphs in different drawing styles. The main objective is to display the data in a meaningful fashion, that is, in a way that visualizes individual nodes and the relationships between them.

Visualization by the tree format reflects the structural classification given by the scaffold tree and allows the user to create a mental map of the hierarchical structure. To ease the construction of the layout and combine the multiple trees ending in the one ring scaffolds to one big tree, a virtual root was added, which is invisible in the final drawing. This is strictly necessary to allow visualization by state-of-the art tree drawing algorithms.

The interactive navigation through the graph poses a challenge to graph drawing algorithms. Whereas traditional graph drawing address the visualization of static graphs, the scaffold trees change over time, and the layout has to be adjusted in real time with as little change as possible.

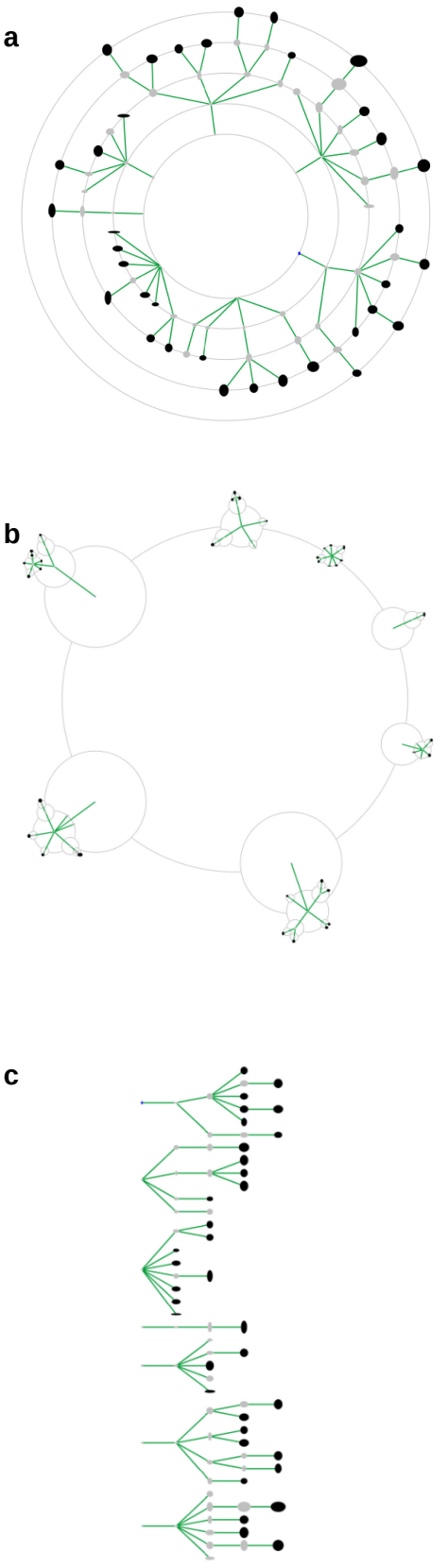


Figure 10: Tree layouts implemented in Scaffold Hunter: a) radial layout, b) balloon layout, c) linear layout. The tree shown is always the same.

Three different layout algorithms were implemented in Scaffold Hunter; one of them is a simple linear placement of the scaffolds, whereas the other two are more sophisticated algorithms. The first, “Radial Layout” (see Figure 10a) is the implementation of a layout approach that was first described by Eades[66] and is still one of the best tree layout patterns available. This standard layout for tree diagrams had been used before in the manually drawn image for the publication (see Figure 5) and the radial layout had proven to be intuitively amenable to chemists and biologists.

Here, the root is placed in the centre of the drawing. The space around this centre node is divided into angular sectors, like slices of pie, for each subtree and objects on the same tree level are placed on concentric circles around the centre. In brief, the algorithm calculates the fraction of the circumference per node from the number of nodes on the outermost circle. Then every branch gets a pie chart according to the number of nodes on the outmost circle multiplied with the fraction of the circumference per node. The nodes are then evenly distributed within the resulting pie chart. The challenge when charting chemical space were the different sizes of the scaffolds. Eades assumed nodes of equal size and, hence, used a static radius for each hierarchy level in his algorithm. Scaffold structures, however, tend to differ in size and, therefore, the radius of each hierarchy level needs to be calculated dynamically during zooming to avoid overlap of scaffold structures.

The second graph layout, the so called “Balloon Layout” (see Figure 10b), follows closely the descriptions from Carriere and Kazman[67] and Lin and Yen[68]. In brief, each subtree is entirely enclosed in a circle which resides in a pie slice whose endpoint is the parent node of the subtree. While the radial layout supports the impression of a tree hierarchy better, the balloon layout makes better use of the available space and allows more focused real time interaction with the user. It may also be more suitable for closer inspection of individual branches in a tree since more hierarchy levels can be visualized at the same time in a space efficient manner.

A linear layout (see Figure 10c) was also implemented to allow for sorting of branches according to scaffold similarity on one hierarchy level. Implementing this feature in a radial layout would have resulted in the most dissimilar scaffolds ending up next to each other on the closed circle of one hierarchy level. In a linear layout, the ordering from top to bottom according to descending similarity is more intuitive.

The user may choose the layout algorithm that is most applicable to his personal preferences or to the task-specific needs. The change of layout is possible while working with Scaffold Hunter and does not need a restart of the program.

User Interface

To facilitate exploration of the data, Scaffold Hunter supports a process of iterative filtering and selection that exploits the user’s domain expert knowledge in combination with the

computational power of a computer. Due to the large amounts of data generated by the high throughput methods of modern biomedical research, filtering the data set according to several parameters is an initial key step for any subsequent analysis, especially for visual analysis of the data. Therefore, a number of mechanisms were implemented to filter data in different ways, for instance by automatic filtering or by colour-coding and subsequent drawing of sub trees comprising of selected scaffolds. These mechanisms will be described in the following paragraphs together with other functionality to enable easy and intuitive navigation in the data.

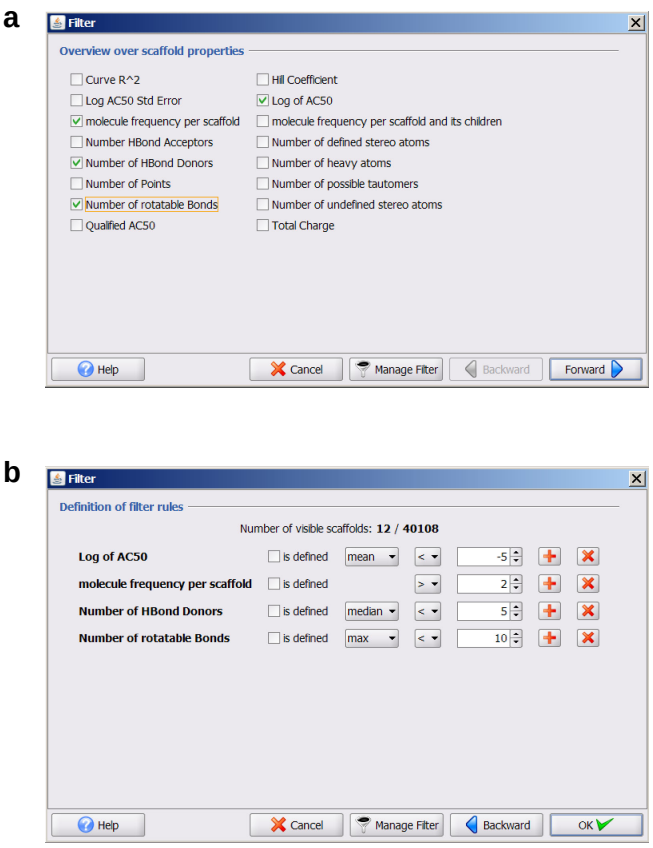


Figure 11: The filter dialogue of Scaffold Hunter is implemented as a two step process: first the user selects the criteria to be used as shown in the upper image. The choice of the statistical measure as well as the corresponding values is then done in a second step as depicted in the lower image.

On startup, a register dialog opens that allows the user to load his personal profile from the database containing global information such as the preferred language (Scaffold Hunter is fully translated into English and German), database connection and layout option settings. A connection to the database is then established to retrieve the set of scaffold properties and their minimum and maximum values from the database. A filter dialog opens where an initial set of properties to be invoked into the filter can be chosen (see Figure 11a). The next window shows these properties and allows the selection of their calculation method, e.g., mean, median, standard deviation etc., and of the filtering bounds for the selected properties (see Figure 11b). The scaffolds that conform to these filter rules are then used to populate the views of the application window, i.e., the scaffold tree that is opened upon confirmation of the filter rules by the user. The number of scaffolds that are initially shown can be further reduced by restricting the number of tree levels that are displayed upon start.

The application window is divided into two main parts: The canvas window on the right and a vertical information bar on the left, see Figure 12. The canvas window shows the main graph view where the user can directly interact with the graphical depiction of the data, explore the

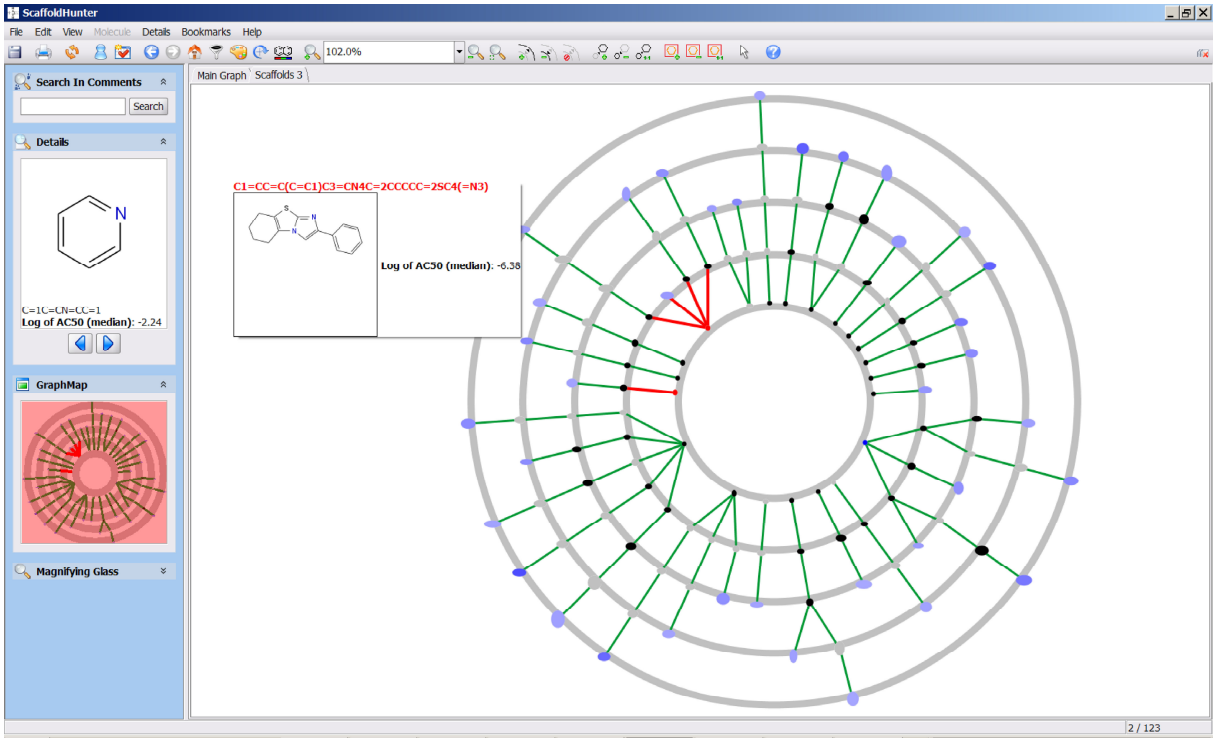


Figure 12: Scaffold Hunter user interface showing the filtered pyruvate kinase scaffold tree at a higher level of the semantic zoom. A scaffold structure can be seen in the tool tip window. On the left hand side, one can see the details window and the graph map.

graph and open new views that show selected sub-graphs. Each scaffold in the database whose properties fall within the ranges defined by the applied filter is represented by a node in the graph. At a higher zoom level, the so called ‘semantic zoom’ shows only solid circles at the nodes instead of the structures that would not be in any case recognizable at this zoom level. Scaffold structures are shown by a mouse-over window that pops up when the mouse cursor is moved onto a node and remains there (see Figure 12).

The information bar on the left hand side of the screen comprises a zoom-and-pan window and a ‘details view’ that shows structure and properties of the selected scaffold(s). If needed, additional windows can be added there, e.g. legends for multi-colour filters etc. The selection of properties that are shown in the Details window and the tool tip can be adjusted to the user’s needs in the options dialog. A toolbar allows direct access to the main features like colouring, filtering and export dialogs. The static presentation of the data, for example for presentations or publications, is supported by picture export and printing capabilities in multiple formats.

The information that is presented and also the visualization can be adapted to the user's needs by the filtering, colouring and selecting of different layout styles. The order in which the scaffolds are arranged around their parent is currently implemented to be computed randomly, but any user-defined sorting routine can be used there instead.

The implemented filters (see Figure 11) are quite generic to allow filtering by all numerical properties stored in the database. One can choose multiple criteria at the same time, which will

be automatically combined by the program. Maximum and minimum values for each filter criterion are determined from the database and limit the range of the filter. The number of visible scaffolds resulting from the actual filter settings is displayed to allow a quick assessment of the criteria applied. If the filter is too wide, i.e., too many scaffolds are selected for display, Scaffold Hunter issues a warning to the user. The limit scaffold number for the warning can be set in the user profile and depends on the size of memory dedicated to the Java Runtime environment when running Scaffold Hunter. According to experience, up to 1,500 scaffolds can be shown at a dedicated memory size of 1024 MBytes. This may, however, depend on the other programs running at the same time and the operating system.

Navigation and Interaction

The automated graph layout provides a static picture of the scaffold hierarchy. The Scaffold Hunter interface allows interactive selection of data to be displayed and animated graph navigation to facilitate exploration of the underlying data. The user can expand and collapse parts of the graph during exploration to hide parts of minor interest for the current task and focus on the relevant and most promising parts for further exploration. The dynamic update of the graph layout and the data views is done in real time and the change in the graph layout is animated smoothly allowing the user to keep his mental map of the data representation. The navigation, for example, moving around, zooming, selecting and expanding, can be done using the mouse or keyboard or a combination of both.

The user can zoom into the canvas view to get a detailed view of the branches and sub branches of interest. A semantic zoom feature that changes the representation of the scaffolds from an abstract rectangle depiction to the fully rendered scaffold structure SVG image allows the display of up to several thousand objects at the same time on lower zoom levels without using too many computational resources on the display of small, unreadable structures. At higher zoom levels, the view switches to the detailed scaffold structures (see Figure 15). As discussed in the layout part of this section, the distance between the concentric circles, i.e., the hierarchy levels, is automatically adapted according to the zoom level. This effectively avoids the display of only a single scaffold with a lot of of empty space around it on lower zoom levels as well as overlay of scaffold structures on lower zoom levels. The distance between the radii can be changed manually as well as the size of selected scaffolds. This feature can be used to highlight selected scaffolds in the tree and provide a quick, intuitive overview of the scaffolds that passed the applied filters (see Figure 13).

Individual branches as well as scaffold trees comprising of selected scaffolds can be opened and viewed in a new tab (see Figure 16). It is also possible to open a full branch in a new tab – irrespective of the initial set of filters applied to the main graph. These features enable quick visualization of the data of interest selected from the large set. Moreover, individual branches

with interesting molecular structures and properties can be studied in more detail.

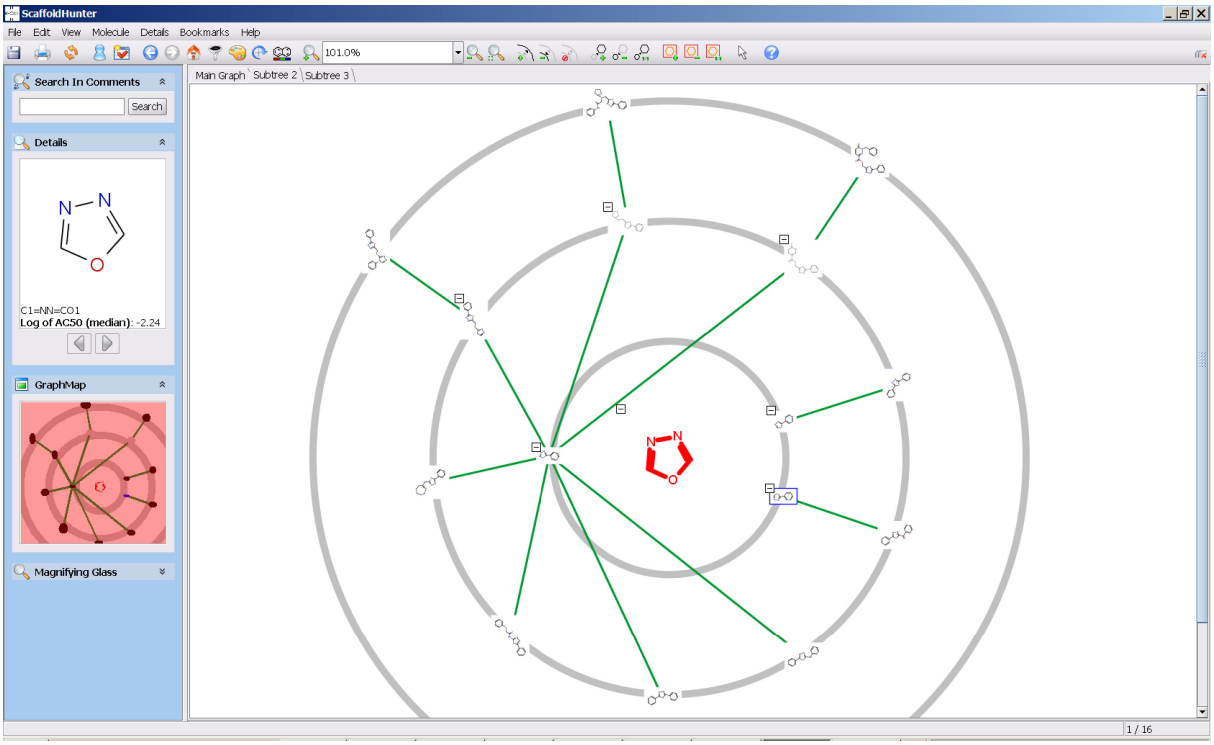


Figure 13: Scaffold Hunter screenshot showing a subtree representing one branch in the large scaffold tree. The root can be seen in the center while the branches stretch out towards the edges of the window.

The user can browse through the selected scaffolds by using the arrow buttons in the Details view. Within the view, the structure SVG of the currently selected scaffold is rendered and a number of properties are shown which can be freely selected by the user from the set of all available properties in the database. A mouse-over detail pop-up shows the same information for non-selected objects when moving the mouse over them. The selected scaffolds are also added to a list in the 'Details' menu for quick access. A click on a scaffold in the Details window zooms in on the position of the scaffold in the tree.

To ensure optimal orientation, especially in larger scaffold trees, the overview-and-detail paradigm was implemented by using a standard zoom-and-pan approach. The graph map shows an overview depiction of the graph consisting of the scaffolds that conform to the filter rules and are currently expanded. The selection of the scaffold tree shown in the main window can directly be chosen by clicking and dragging open a rectangle in the graph map.

To facilitate comparison of scaffolds depending on their properties colour coding and shading features were implemented in Scaffold Hunter. Scaffolds can be shaded depending on the value of a specific property, i.e., scaffolds with a higher value get darker shading than scaffolds with lower values (see Figure 14). Scaffolds can also be coloured by defining a minimum and maximum value of one or more properties which will colour only those scaffolds that conform to these boundaries. Colouring can also be done based on the definition of a filter set whose result

set will then be coloured accordingly. This colour-coding of underlying, even complex information greatly facilitates data mining, especially in data sets with multiple annotations.

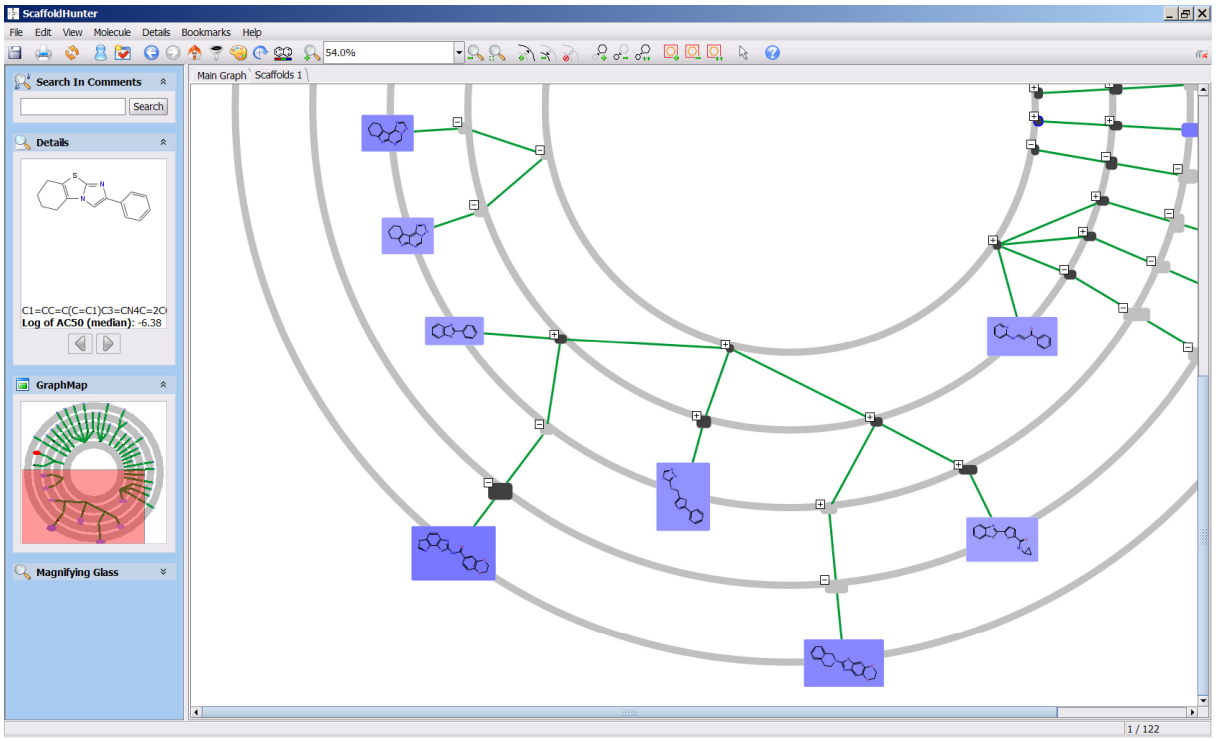


Figure 14: Scaffold Hunter screenshot showing a subtree with nodes coloured according to the median activity of the molecules represented by each scaffold. The darker the colour the more active are the molecules represented by that particular scaffold on average.€

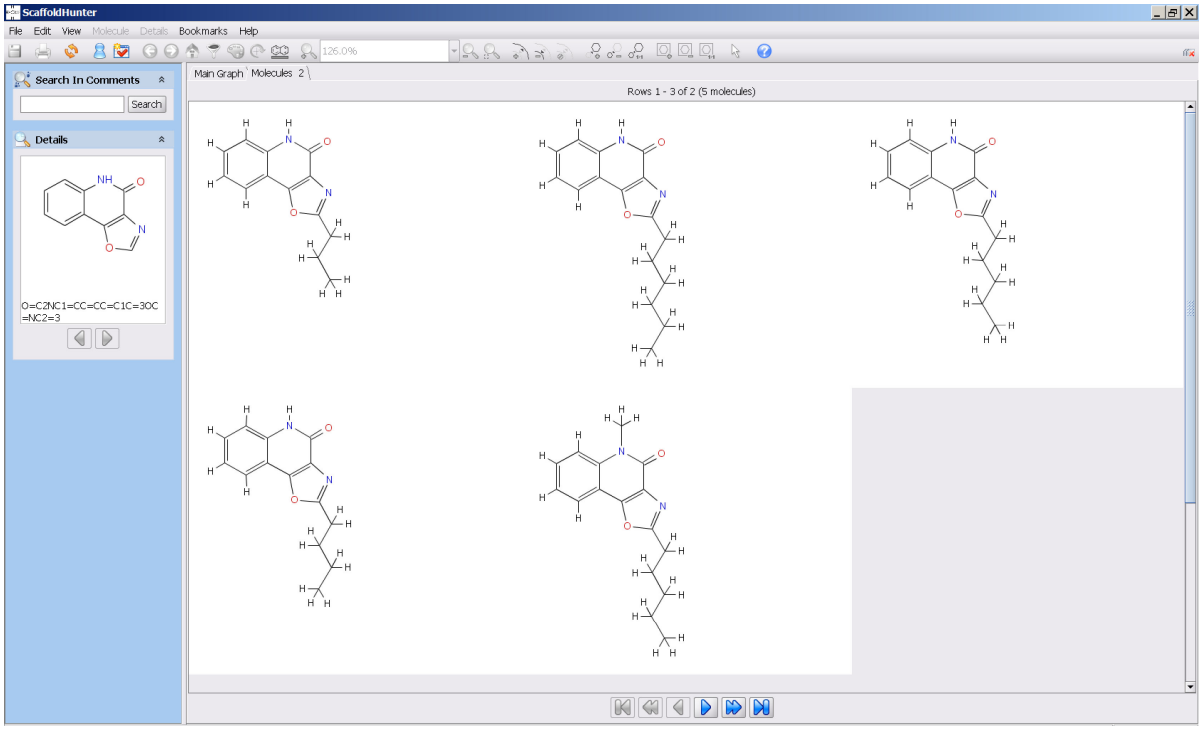


Figure 15: View on the molecules represented by a scaffold.

A one-click function that selects all coloured scaffolds in combination with the “selected scaffold in new tab” function from the context menu facilitates the generation of a focused sub-tree of selected scaffolds with two mouse clicks. Thus, a set of scaffolds or a subtree of the hierarchy selected for a closer and more careful inspection can easily be displayed in a new view and be analyzed. The new tab offers the full functionality of the main window and allows full navigation through the data.

A set of scaffolds that the user wants to mark for further investigation, e.g., in later sessions, can be added to a bookmark list.

The structures of the molecules represented by each scaffold are accessible via a double click on the corresponding scaffold structure (see Figure 15). The scaffold structure representing the molecules is shown in the Details window.

1.3.5 Exploring gaps in chemical space: novel kinase inhibitor chemotypes

A particularly interesting feature of Scaffold Hunter is that the program identifies white spots on the chemical space map, which represent virtual scaffolds. Virtual Scaffolds are scaffolds that result from the ring pruning procedure but do not represent existing scaffolds of molecules in the dataset. Notably, "brachiation" along branches of the scaffold tree from larger more complex scaffolds towards smaller scaffolds by analogy to locomotion of primates along branches in botanical trees may lead to the identification of less complex compound classes with retained biochemical activity.[23,44] Such "brachiation" along lines of biological relevance differs from the generally accepted "similarity principle", i.e., that similar molecules likely share similar properties.[69] Rather it is based on the assumption that smaller scaffolds incorporated into larger molecules are likely to share some of their properties. A similar reasoning also underlies current fragment-based screening and design approaches where fragments with low affinity are linked to form molecules with higher affinity.[70-73] In line with the brachiation approach, virtual scaffolds can be expected to share properties with their parent or child compounds, in particular their bioactivity. The set of rules guiding the ring pruning procedure was designed to render virtual scaffolds chemically meaningful entities when derived from a known molecule. Hence,

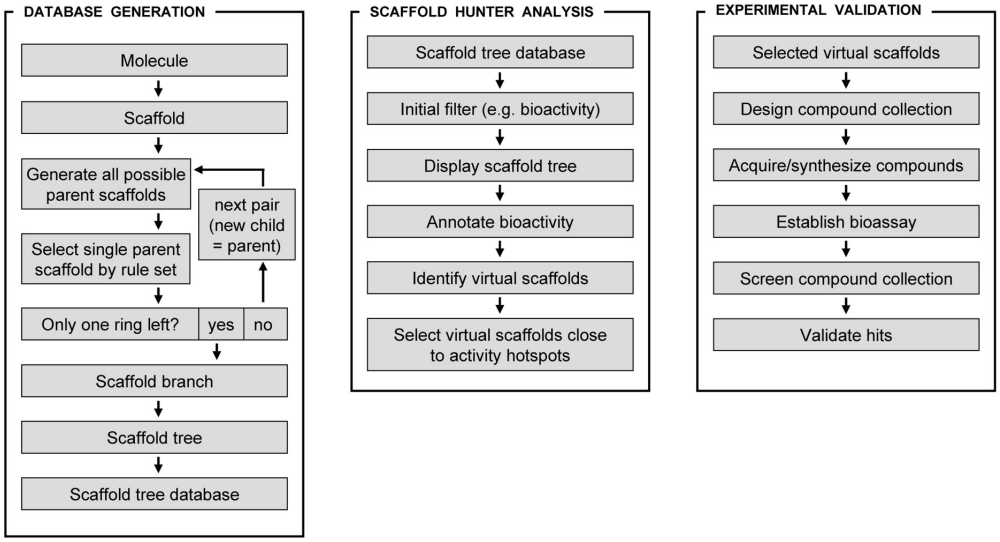


Figure 16: Visualization of the scaffold tree database generation process. Firstly, the molecule is reduced to the chemically meaningful scaffold from which all possible parent scaffolds with one ring less are generated. Following the set of rules, one parent scaffold is selected to form the parent-child pair. If this parent scaffold can be deconstructed further, the next pair is formed using the generated parent scaffold as new child. Once the scaffold has been fully deconstructed, the resulting branch is generated. Combination of all branches from a data set forms the scaffold tree that is written to a database. Exploration of gaps in chemical space starts with a Scaffold Hunter analysis for promising virtual scaffolds. After startup, Scaffold Hunter reads those scaffolds from the database that pass the initial user-defined filter and displays the scaffold tree formed by them. Annotation of bioactivity, e.g. by colour shading, together with virtual scaffolds shown in grey facilitates the quick identification of promising virtual scaffolds close to activity hotspots. Compounds incorporating these scaffolds are then acquired or synthesized and tested in a biochemical assay for the target of interest.

virtual scaffolds may provide new opportunities for the identification of novel biologically relevant scaffold classes. This process is illustrated in Figure 16.

In order to investigate whether the guiding suggestion provided by identified white spots in complex sub fractions of chemical space can indeed lead to the identification of new active compound classes, a first analysis of HTS data available from PubChem was performed. Subsequently, the known literature data, as contained in WOMBAT, was searched for compounds filling the gaps, i.e., incorporating a PubChem virtual scaffold and exhibiting biochemical activity on the same molecular target. For data analysis the scaffold tree of the 765,135 ring-containing structures was generated, for which biological or biochemical screening data are available in PubChem. These structures were annotated with the corresponding results from confirmatory concentration-dependent measurements. To match the screens with the bioactivity data contained in the WOMBAT database[74], the molecular target of each screen was annotated with its UniProt ID[75]. These efforts yielded 60 different targets present in both PubChem and in the WOMBAT database, corresponding to 199 individual assays in PubChem. Of these 199 assays, in 102 cases concentration-dependent measurements were carried out and IC_{50}, EC_{50} or GI_{50} values were calculated for 46 different targets. Further analysis of the scaffold tree comprising of all compounds screened on a particular target (at one fixed concentration as well as concentration-dependent measurements) identified the scaffolds representing the active compounds as defined by PubChem. In the next step, the virtual parents of these compounds, i.e., smaller scaffolds that are not represented by compounds for which screening data are available against this particular target in PubChem, were selected. A subsequent WOMBAT query for compounds filling the gaps identified from PubChem data, i.e., compounds annotated with bioactivity for the target of interest and embodying the virtual scaffold substructure, yielded 517 active compounds. These represent 14 virtual scaffolds active against 5 different targets (see Table 2).

These results indicate that Scaffold Hunter generates virtual scaffolds that are likely to have biological or biochemical relevance and, thus, represent valuable compounds to be acquired and screened.

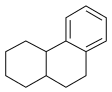
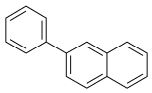
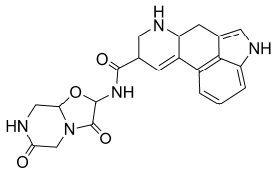
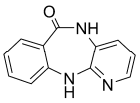
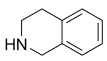
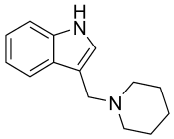
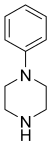
Entry	Target (Swiss Prot ID)	Virtual scaffold from screen in PubChem	Number of scaffold families found to be active in WOMBAT	Number of compounds represented by these scaffolds
1	Estrogen Receptor-alpha (P03372)		9	169
2	Estrogen Receptor-alpha (P03372)		6	74
3	5-HT Receptor Subtype 1A (P08908)		4	9
4	5-HT Receptor Subtype 1A (P08908)		1	2
5	Acetylcholine Muscarinic M1 Receptor (P11229)		1	1
6	Acetylcholine Muscarinic M1 Receptor (P11229)		1	1
7	Acetylcholine Muscarinic M1 Receptor (P11229)		54	131
8	Acetylcholine Muscarinic M1 Receptor (P11229)		12	25
9	Acetylcholine Muscarinic M1 Receptor (P11229)		1	1
10	Acetylcholine Muscarinic M1 Receptor (P11229)		9	98
11	Acetylcholine Muscarinic M1 Receptor (P11229)		1	1
12	Thyroid Stimulating Hormone Receptor (P16473)		1	1
13	Allosteric Modulators of D1 Receptors (P21918)		1	1
14	Allosteric Modulators of D1 Receptors (P21918)		2	2

Table 2: Data for virtual scaffolds from PubChem present in compounds described as active against the same molecular target in the WOMBAT database.

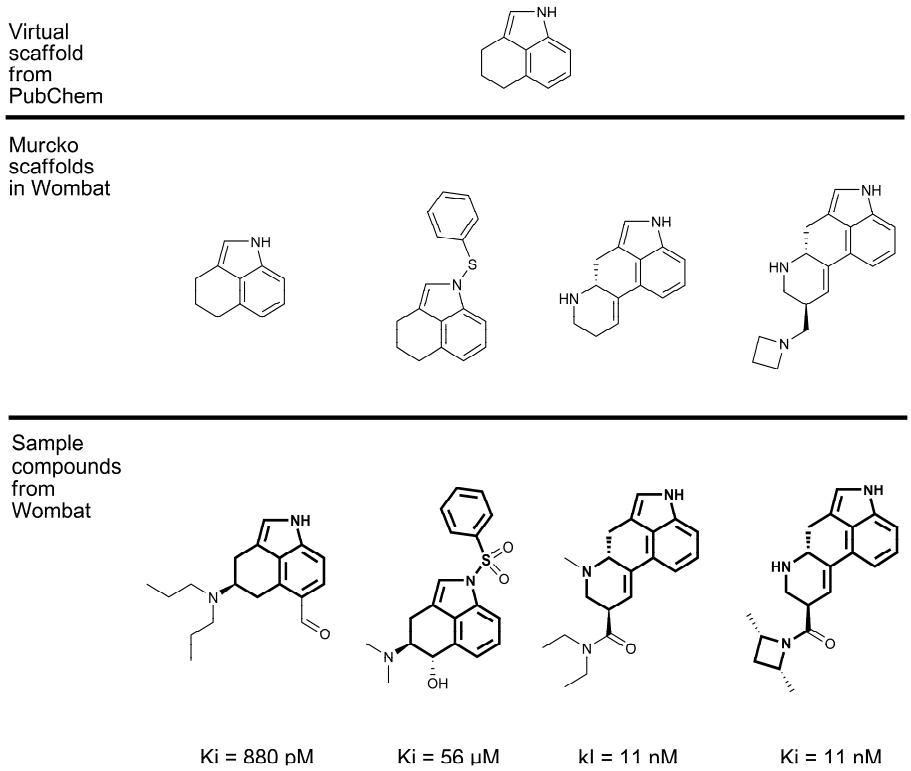


Figure 17: Representative example for a virtual scaffold from a Serotonin 5-HT_{1a} receptor screen in PubChem, the Murcko families and their corresponding compounds found in the WOMBAT search.

Prospective exploration of gaps in chemical space predicted from sets of biochemical data

To further validate the target annotation with Scaffold Hunter in a prospective manner, a set of HTS data was analyzed in order to predict promising virtual scaffolds and acquire and experimentally validate the activity of compounds incorporating the predicted scaffolds. A concentration dependent screening data set available from PubChem was chosen that contains 51,415 unique molecules tested as inhibitors or activators of pyruvate kinase.[76,77] The original report identified 602 biochemically active compounds of which 472 are inhibitors and 130 are activators.[76] A compound was labeled "active" if the AC_{50} was smaller than or equal to 10 µm. From the structures of the compounds included in the screen, the scaffold tree algorithm generated a total of 35,868 scaffolds distributed over 767 branches. Of these, 24% (8,684) were identified as "virtual scaffolds". Of the remaining scaffolds, 385 represent at least one inhibitor and 123 represent at least one activator.

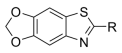
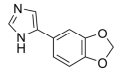
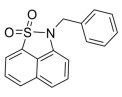
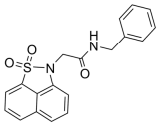
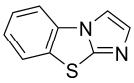
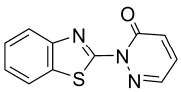
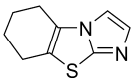
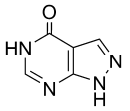
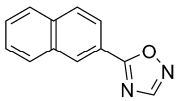
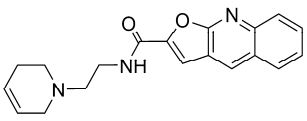
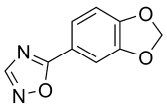
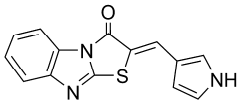
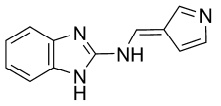
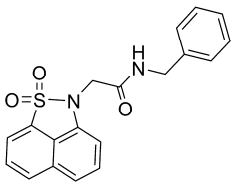
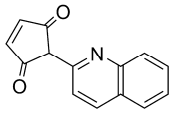
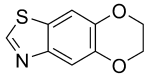
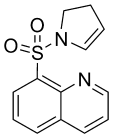
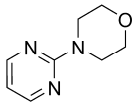
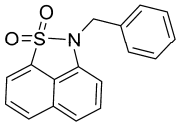
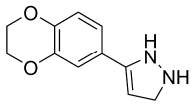
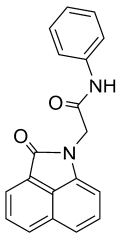
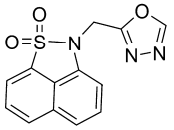
Num.	Virtual scaffold	Compounds purchased
1		21
2		57
3		14
4		15

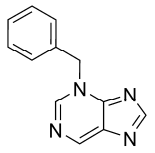
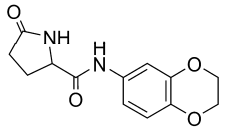
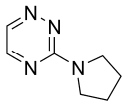
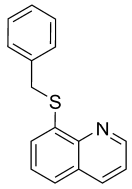
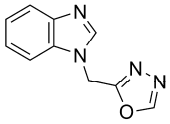
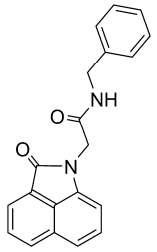
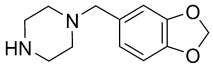
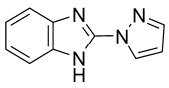
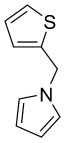
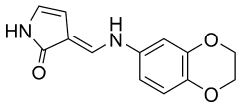
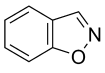
Table 3: Selected virtual scaffolds identified as promising structural templates for modulators of pyruvate kinase activity. The number of compounds acquired and tested for each scaffold is also provided.

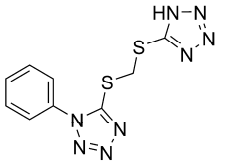
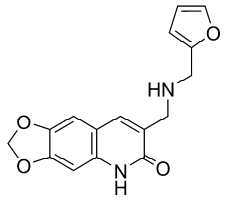
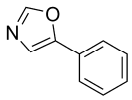
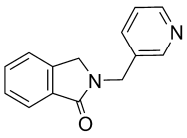
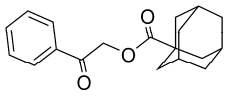
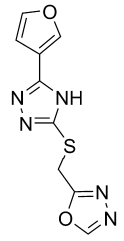
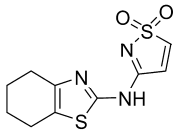
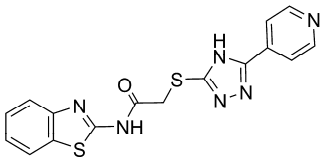
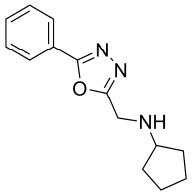
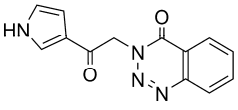
Filtering for scaffolds with mean AC_{50}[76] values lower than 100 µM reduced the number of scaffolds to 1800. Subsequent selection for AC_{50} values smaller than 10 µM yielded a focused scaffold tree consisting of only the selected scaffolds and their branches to the root scaffold. Color-shading of the tree according to the AC_{50} value highlighted the active scaffolds and enabled the identification of virtual scaffolds that are close neighbours to the scaffolds of active molecules (Figure 18). This highly interactive and intuitive analysis is demonstrated in Supporting Movie 1 that is available from the Nature Chemical Biology website: http://www.nature.com/nchembio/journal/vaop/ncurrent/suppinfo/nchembio.187_S1.html.

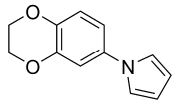
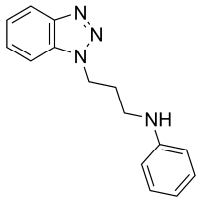
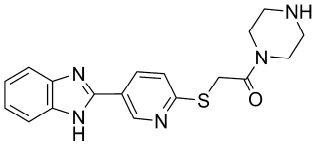
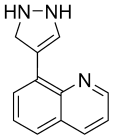
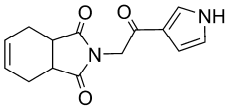
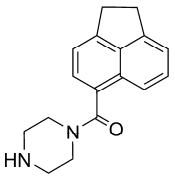
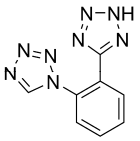
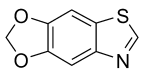
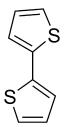
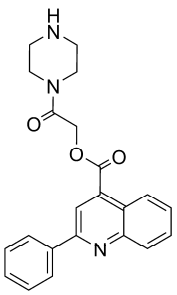
The analysis yielded 65 virtual scaffolds related to scaffolds with good activity (mean log AC_{50} < -5.00 (equalling a mean AC_{50} < 100 µM), for all virtual scaffolds see Table 4). Four of these 65 scaffolds were selected (the branches to which they are assigned are partially shown in Figure 18 a,b,c, see also Table 3) and 107 compounds embodying these virtual scaffolds as substructure were acquired from commercial sources (see Attachment 1).[78,79] The compounds were screened in a pyruvate kinase assay under conditions similar to those originally published.[77,80] Compounds identified as either potent inhibitor or activator of pyruvate kinase in an initial pre-screen were subjected to concentration dependent measurements to determine IC_{50} values (Table 5).

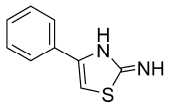
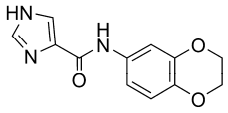
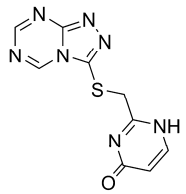
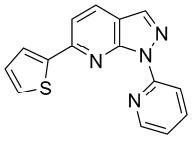
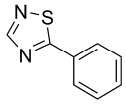
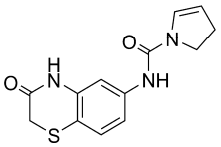
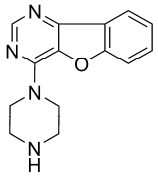
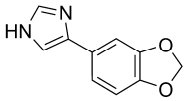
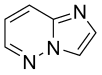
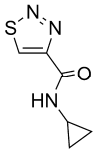
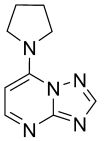
Number	Virtual scaffold structure	Scaffold hierarchy level	Avg. log AC50 of parent scaffold(s)	Avg. log AC50 of child scaffold(s)
1		3	-2.47	-7.26
2		3	-3.06	-6.79
3		3	-2.47	-6.38
4		2	-2.29	-6.27
5		3	-3.20	-6.13
6		4	-3.64	-6.12
7		3	-2.34	-6.07

Number	Virtual scaffold structure	Scaffold hierarchy level	Avg. log AC50 of parent scaffold(s)	Avg. log AC50 of child scaffold(s)
8		4	-3.84	-6.07
9		3	-2.37	-6.05
10		4	-5.02	-6.03
11		3	-2.99	-5.94
12		3	-3.02	-5.90
13		3	-2.99	-5.84
14		2	-2.36	-5.83
15		4	-5.02	-5.82
16		3	-2.32	-5.73
17		4	-2.79	-5.72
18		4	-5.02	-5.68

Number	Virtual scaffold structure	Scaffold hierarchy level	Avg. log AC50 of parent scaffold(s)	Avg. log AC50 of child scaffold(s)
19		3	-2.24	-5.66
20		3	-2.32	-5.64
21		2	-2.45	-5.62
22		3	-2.99	-5.61
23		3	-2.37	-5.60
24		4	-2.79	-5.56
25		3	-2.34	-5.55
26		3	-2.37	-5.55
27		2	-2.31	-5.50
28		3	-2.32	-5.50
29		2	-2.27	-5.47

Number	Virtual scaffold structure	Scaffold hierarchy level	Avg. log AC50 of parent scaffold(s)	Avg. log AC50 of child scaffold(s)
30		3	-2.51	-5.46
31		4	-2.24	-5.42
32		2	-2.24	-5.40
33		3	-2.24	-5.38
34		2	-2.26	-5.37
35		3	-2.24	-5.35
36		3	-2.59	-5.35
37		4	-2.24	-5.33
38		3	-2.75	-5.33
39		3	-2.24	-5.25

Number	Virtual scaffold structure	Scaffold hierarchy level	Avg. log AC50 of parent scaffold(s)	Avg. log AC50 of child scaffold(s)
40		3	-2.32	-5.25
41		3	-2.24	-5.24
42		4	-3.16	-5.24
43		3	-2.99	-5.23
44		3	-2.24	-5.22
45		4	-2.59	-5.22
46		3	-2.52	-5.21
47		3	-3.06	-5.21
48		2	-2.40	-5.20
49		4	-3.79	-5.19

Number	Virtual scaffold structure	Scaffold hierarchy level	Avg. log AC50 of parent scaffold(s)	Avg. log AC50 of child scaffold(s)
50		2	-2.24	-5.16
51		3	-2.32	-5.16
52		3	-2.94	-5.16
53		4	-2.24	-5.15
54		2	-2.24	-5.11
55		3	-2.29	-5.09
56		4	-3.39	-5.07
57		3	-2.34	-5.05
58		3	-2.42	-5.05
59		2	-2.24	-5.04
60		3	-2.24	-5.04

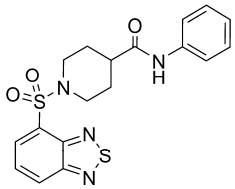
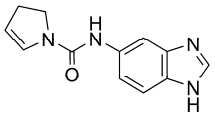
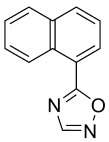
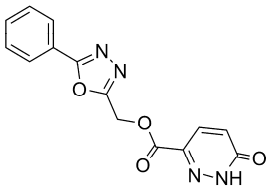
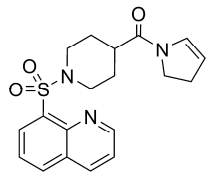
Number	Virtual scaffold structure	Scaffold hierarchy level	Avg. log AC50 of parent scaffold(s)	Avg. log AC50 of child scaffold(s)
61		4	-4.30	-5.03
62		3	-2.37	-5.02
63		3	-3.21	-5.02
64		3	-2.75	-5.01
65		4	-4.76	-5.01

Table 4: virtual scaffolds that are neighbours to scaffolds representing active small molecule modulators of pyruvate kinase activity.

The screen identified nine compounds that showed activity as pyruvate kinase inhibitors and activators with AC_{50} values in the 1-10 µM range (Table 5). So far, none of these compounds nor their any of their scaffolds had been described as inhibitors or activators of pyruvate kinase according to a search in Chemical Abstracts using SciFinder.[81] As expected, the activity type determined for these compounds matched the activity of the compounds in the guiding tree branches, i.e., virtual scaffolds from branches with inhibitors yielded inhibitors and branches with activators yielded activators. Notably, brachiation from the scaffolds stemming from the original data set to the newly identified inhibitors included more than one brachiation step, thereby efficiently guiding potential synthesis efforts.

To further characterize the hits with regard to their medicinal chemistry properties, their compliance with the Rule-of-Five[3] and the lead-likeness criteria derived by Verheij *et al*. was investigated.[82] The Rule-of-Five defines molecular parameters empirically derived from a set of orally bioavailable drugs by Lipinski *et al*.[3] Compliance with the Rule-of-Five often is a requirement for development candidates in the pharmaceutical industry. The lead-likeness

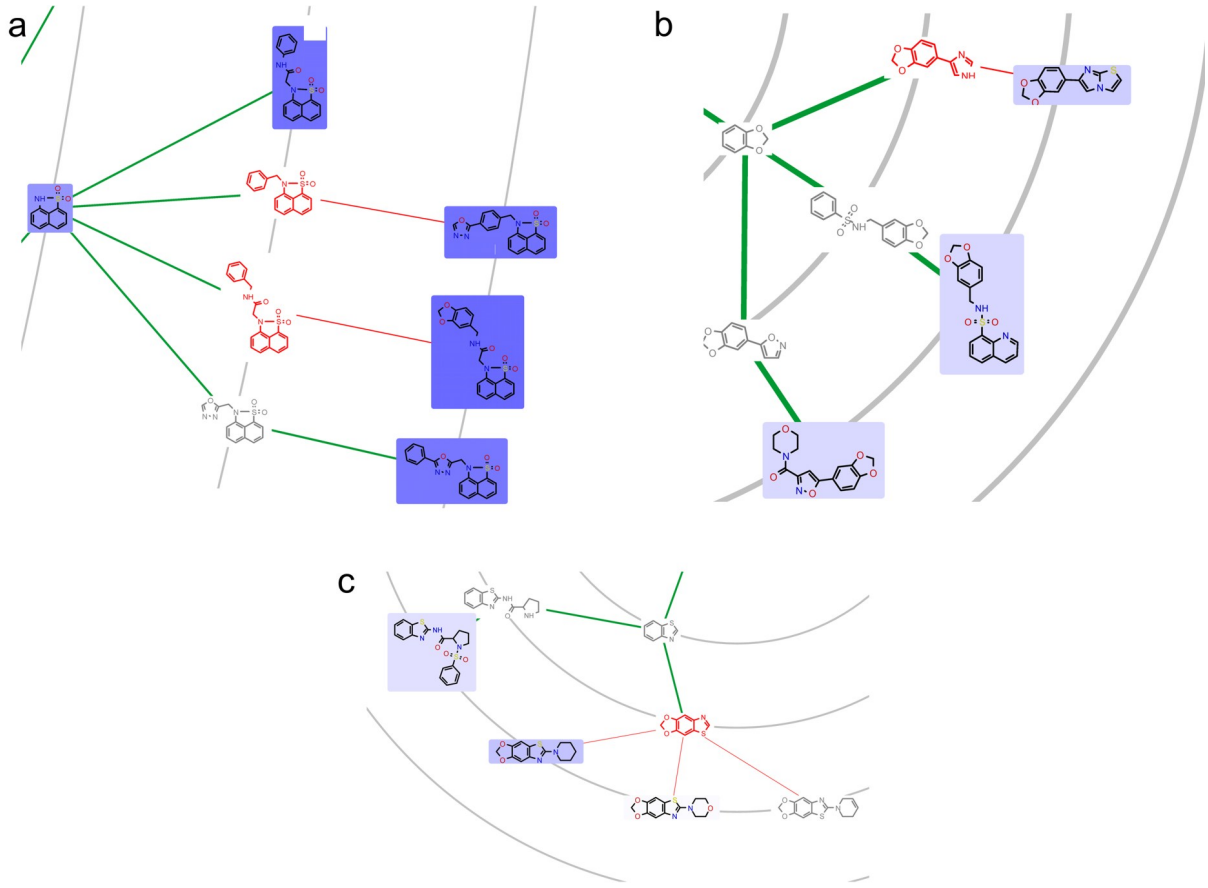


Figure 18: Branches of the scaffold tree derived from the pyruvate kinase HTS dataset. The four virtual scaffolds selected for compound acquisition are shown in red. a) This branch consists of several scaffolds representing good activators of pyruvate kinase from which two virtual scaffolds were picked for compound acquisition. b), c) Branches which represent inhibitors of pyruvate kinase and from which virtual scaffolds were chosen for compound acquisition. Additional virtual scaffolds are shown in grey. Blue colour-shading highlights the mean log AC_{50} values obtained from the dataset (darker shading represents higher activity).

criteria are based on the empirical observation that chemical modifications of lead compounds in late stages of lead development mainly increase molecular weight and hydrophobicity.[83,84] Therefore, lead-like compounds are generally characterizied by smaller size and higher polarity which is well in agreement with the brachiation principle of generating small active molecules that are part of larger active child molecules. All of the hits discovered by this approach fulfil the criteria of the Rule-of-Five as well as the lead-likeness criteria. In both sets, almost all hits satisfy the Rule-of-Five. The criteria for lead-likeness according to Verheij *et al*.[82] are met by 86% of the PubChem and by all compounds identified by the screen. To compare these hits to the hits found in PubChem, the ligand-binding efficiency[85] was calculated, i.e., the binding affinity per heavy atom. This measure reflects the general trends that a) small potent compounds are better suited for further development and b) unspecific binding increases with molecular weight. The analysis shows that the ligand-binding efficiency of the compounds discovered with Scaffold Hunter and the hits discovered by the screen published in PubChem compare well (see Figure 19).

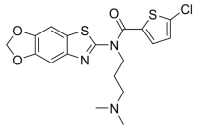
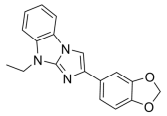
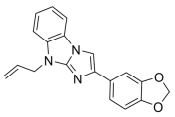
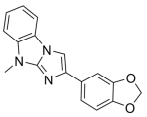
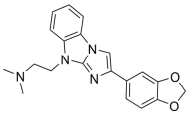
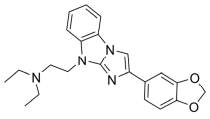
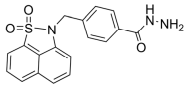
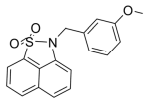
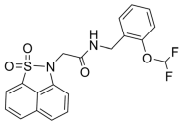
Num.	Structure	Activity type	AC_{50} [µM]
PKL-39		Inhibitor	8.9 ± 1.1
PKL-66		Inhibitor	2.0 ± 0.3
PKL-68		Inhibitor	10.5 ± 1.6
PKL-65		Inhibitor	1.0 ± 0.1
PKL-69		Inhibitor	1.7 ± 0.3
PKL-70		Inhibitor	2.2 ± 0.3
PKL-1		Activator	4.9 ± 0.6
PKL-7		Activator	5.4 ± 0.6
PKL-21		Activator	7.6 ± 0.6

Table 5: Hits discovered in the pyruvate kinase screen of the selected virtual scaffold libraries. All compounds of the pyruvate kinase library (PKL) were sequentially numbered as shown in Attachment 1.

Finally, as a negative control it was investigated whether compound collections based on virtual scaffolds from branches mainly representing inactive compounds also show little or no activity. To his end, 88 compounds were selected and tested (see Attachment 2) from branches where parent and child scaffolds represent mainly inactive compounds, i.e., they exhibit a mean log

AC_{50} > -2.5 (average log AC_{50} values have been calculated exclusively from the values stored in PubChem. Inactive compounds in PubChem generally are annotated with log AC_{50} values of -2.5). The 88 compounds selected represent six different scaffolds and are shown in Attachment 2. A fixed-concentration screen conducted at 100 µM compound concentration indeed confirmed that even at this very high concentration in all cases more than 70% of the enzymatic activity remained. Thus, these compounds are weak pyruvate kinase inhibitors at best.

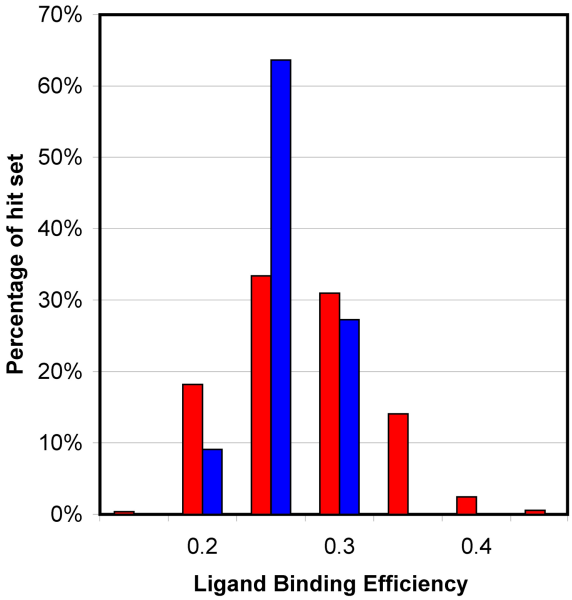


Figure 19: Comparison of the ligand-binding efficiencies of the hits from the PubChem screen (red) and the hits discovered with Scaffold Hunter (blue).

1.3.6 Prospective Bioactivity Annotation by Scaffold Tree Merging

Although natural products have a long standing successful history as source of drug precursors[86], their use in drug discovery programs is often limited by the laborious supply of compound, either through isolation or through total synthesis, and by the lack of information about their biological role and the proteins modulated by them. Whereas the supply problem can be addressed by synthetic protocols aiming at the generation of natural product-derived libraries, e.g. biology oriented synthesis (BIOS)[44,48,87] and diversity oriented synthesis (DOS)[88], target annotation to natural products remains a challenge. Approaches for charting of chemical space can be applied for target annotation by co-mapping a set of natural product structures with a set of reference compounds with known biochemical or biological activity.[17,21] With this method, one should be able to generate a scaffold tree with nodes annotated for biological activity. These annotated notes then form the basis for the target annotation of surrounding scaffolds via brachiation, that is, the simplification of complex molecular scaffolds by movement along the branches of the scaffold tree towards smaller and structurally less complex scaffolds while retaining activity against the target of interest. A scaffold tree analysis of a large data set based on bioactivity as the key criterion for branch construction revealed successful brachiation for many compound and all major pharmaceutical target classes.[89] These findings support the combined co-mapping/brachiation strategy for target annotation outlined above.

In this section, the potential of the scaffold tree approach and Scaffold Hunter for target annotation by co-mapping of synthetic compounds with annotated biological activity and extension of this annotation by subsequent brachiation was evaluated. Then, a number of potential protein targets for the γ-pyrone containing natural product scaffold branch were identified. Experimental testing of a compound collection with 400 γ-pyrones against a target

selection including monoamine oxidase, sphinomyelinase and signal transducer and activator of transcription (STAT) yielded selective inhibitors for each protein class.

Target annotation of natural product chemical space by scaffold tree merging

The natural product chemical space populated by the natural products and analogues contained in the Dictionary of Natural Products (DNP), version 17.2 from 2008[5] was charted by generation of a scaffold tree with the chemistry-based set of rules devised in collaboration with Schuffenhauer and Ertl from Novartis.[56] In brief, compounds were deglycosylated *in silico* and the chemically meaningful scaffold was extracted. Iterative deconstruction one ring at a time generated the scaffold hierarchy forming the branch for this particular chemically meaningful scaffold. All branches were combined to form the natural product scaffold tree that was annotated in the next step with the biological targets of 190,000 molecules contained in the WOMBAT database (release 2007.1). The WOMBAT database lists molecules with proven biochemical or biological activity extracted from literature.[74,90] The previously described co-mapping strategy was used to annotate protein targets to the natural product scaffold tree by merging of the scaffold trees generated from the DNP and WOMBAT (see Figure 20). During this process, identical nodes were fused whereas all other nodes present only in either the DNP or the WOMBAT tree were added to the combined scaffold tree.

The biological activity profiles of compounds in WOMBAT seemed rather well-defined given the fact that more than 75% of the compounds embodying scaffolds with 1-5 rings are described as active against three different proteins or less. Whereas compounds with larger

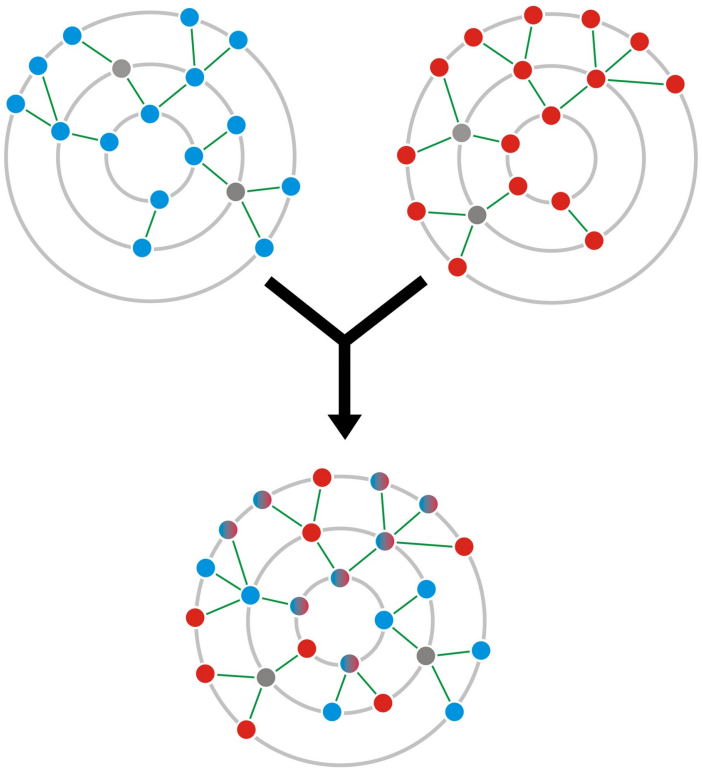


Figure 20: Schematic illustration of the merging of the scaffold trees from the DNP (blue) and Wombat (red). The circles denote the scaffolds in the tree. During the merging, identical scaffolds are mapped onto each other (red/blue circles) while all the other scaffolds are added to the resulting tree. The grey dots denote “virtual scaffolds” that are created by the scaffold tree generation but not chemically meaningful scaffolds of molecules in the data set.

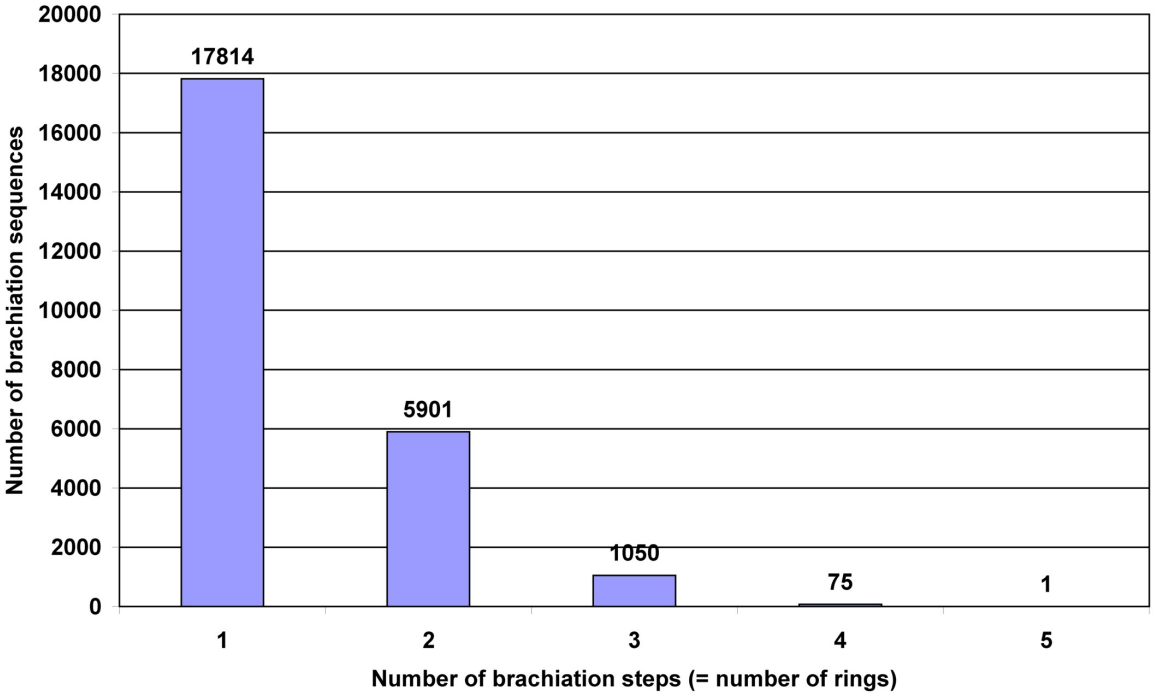


Figure 21: Brachiation sequences without gaps identified from the chemistry-based WOMBAT scaffold tree. For details on the analysis, please refer to the Experimental Section.

scaffolds (6-8 rings) were found to be slightly less selective, the potency distribution showed a clear shift from the micromolar to the nanomolar range for larger scaffolds. This finding correlates well with earlier investigations that revealed highly potent compounds (IC_{50}/K_i < 1nM) as generally more hydrophobic, larger and, as a consequence, less soluble.[91]

To assess the scope of brachiation within the WOMBAT data, the scaffold tree was built according to the chemistry-based set of rules and analyzed for already proven brachiation sequences, i.e. scaffold sub-braches where each scaffold represented compounds active against the same target. Although no gaps were allowed, brachiation sequences with up to four steps, i.e. removal of four (!) rings while retaining biological activity (IC_{50}/K_i ≤ 10 µM) against the same molecular target, were identified. The analysis established 17,814 sequences spanning one brachiation step, 5,901 spanning two, 1,050 with three and 75 with four brachiation steps (see Figure 21). This finding correlated well with the results of Renner *et al*. who identified more brachiation sequences in the WOMBAT data using a scaffold tree constructed according to bioactivity annotation.[89,92]

Analysis of experimentally validated brachiation among γ-pyrones in the natural product tree

To evaluate the scope of the prospective annotation of protein targets via scaffold tree merging, one branch of natural product scaffolds was selected and a compound collection built on scaffolds from this branch was experimentally screened against a selection of annotated targets. Key criteria for the selection of such a branch included a comprehensive coverage of

severals scaffolds in the branch by compounds in the DNP and WOMBAT, annotation of several targets to different nodes of the WOMBAT branch, as well as the availability of a compound library spanning several hierarchy levels of the branch with sufficient substituent diversity. The γ-pyrone branch satisified all of these criteria: it comprised of 8,171 compounds in the DNP and 1,701 in WOMBAT; a significantly large number of compounds to draw conclusions from. WOMBAT linked 133 targets to the γ-pyrone branch whereas more than 95% of the γ-pyrones were described as active against only two targets or less (see Figure 22) which indicated low promiscuity of the compounds. Analysis for documented brachiation sequences in the γ-pyrone branch involving at least three different scaffolds yielded seven examples shown in Table 6.

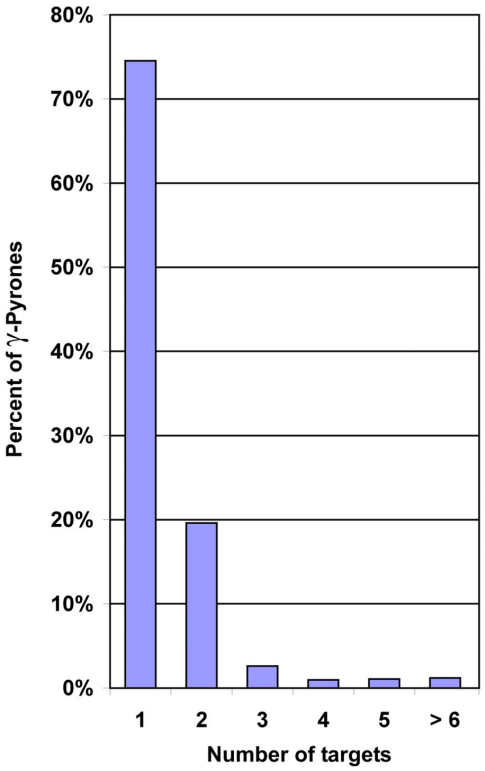
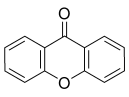
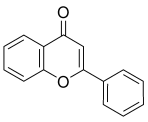
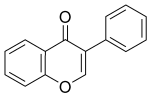


Figure 22: Number of targets annotated in WOMBAT per molecule for the γ-pyrone branch.

No.	Target protein	Swiss Prot ID	Number of inhibitors with IC_{50} < 10 µM per scaffold family				
							
1	Acetylcholinesterase	P22303		3	37	4	
2	Aromatase, estrogen synthetase	P11511		11	7	11	16
3	DNA-dependent protein kinase	P78527	8	131		15	
4	Estrogen receptor α	P11474		4		1	20
5	Estrogen receptor β			3		2	17
6	HIV-1 IN; nucleotidyl-transferase	P03369		2	3	9	
7	steroid sulfatase	P08842		30		1	1
8	acid sphingomyelinase				(3)*		
9	Monoamine oxidase A			2	1		

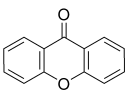
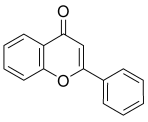
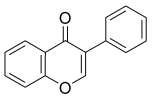
No.	Target protein	Swiss Prot ID	Number of inhibitors with IC_{50} < 10 µM per scaffold family				
							
10	Monoamine oxidase B			5	1		
11	STATs			(15)*			

Table 6: Identified protein targets for γ-pyrone scaffold and distribution of protein activity modulators extracted from WOMBAT over the different γ-pyrone-containing scaffolds. Only targets that are modulated by compounds from at least three different scaffold families in the γ-pyrone branch are shown. Selection criterion: IC_{50} < 10 µM determined in concentration-dependent measurements to ensure a high quality of the data. For the choice of the activity threshold, see Experimental section.

* Number of compounds found active in concentration-dependent manner but IC_{50} ≥ 10µM.

Experimental validation of compounds and targets from the γ-pyrone branch

In numerous previous screens of small focused natural product-derived libraries[23,44] average hit rates of 1-2 % in biochemical assays were observed. This implies that a given library needs to comprise of at least 100 to 200 compounds to attain a reasonable chance of finding hits. In line with these findings, a 500-membered compound collection was assembled comprising of compounds from five different scaffolds from the γ-pyrone branch (see Figure 23, all structures are shown in Attachment 3). The compounds were partially purchased from commercial sources and partially synthesized by Dr. Sammy Chammaa and Dipl.-Chem. Wolfram Wilk. The library was characterised according to the usual parameters including the calculated octanol-water partition coefficient (AlogP[93]), the number of hydrogen bond donors and acceptors and the number of rotatable bonds. The distributions of all parameters were well within the accepted ranges (see Figure 24a) and 78% of the compounds did not violate the Rule-of-Five.[3] Notably, more than 50% of the library inspired by natural-products also satisfied the stricter criteria defined for lead-likeness (see Figure 24b).[31,33,34,83,91]

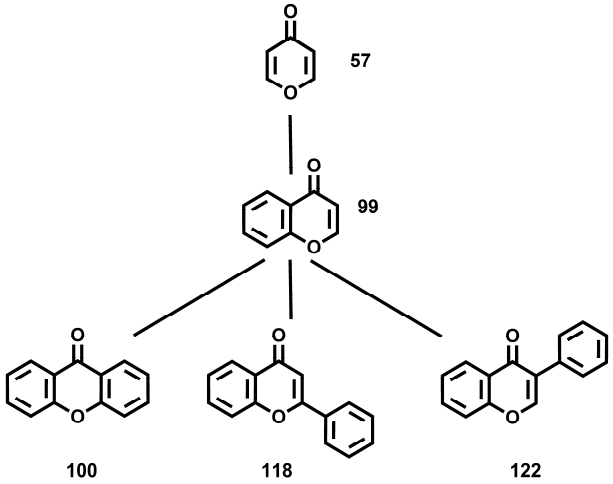


Figure 23: composition of the γ-pyrone library. The five γ-pyrone scaffolds are shown together with the number of compounds embodying each scaffold.

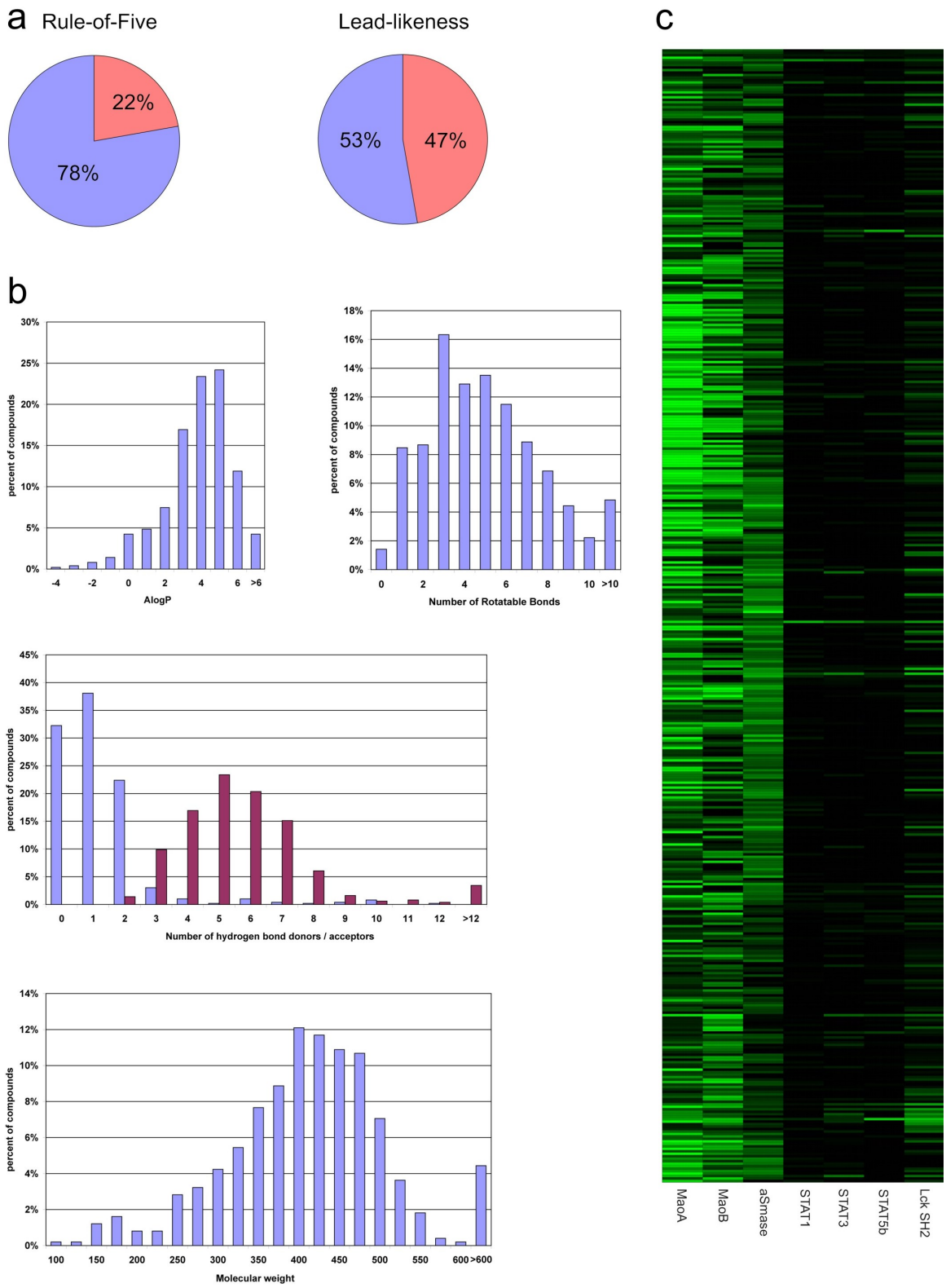


Figure 24: Library statistics and screening results for the 500-membered γ-pyrone library. a) a high percentage of the compounds are drug-like and many are lead-like as well. b) distributions of AlogP, the number of hydrogen bond donors (blue) / acceptors (red), number of rotatable bonds and molecular weight are well within the limits defined, e.g. by the Rule-of-Five. c) Heatmap showing the pre-screen results of the biochemical assays for the γ-pyrone library against Mao A and B, acid Sphingomyelinase (aSmase), STATs 1,3 and 5b as well as the SH2 domain of Lck. Black colour denotes 100% residual activity whereas green colour indicates 0% residual activity (= full inhibition at the given concentration).

From the 133 proteins with annotated activity for compounds embodying the γ-pyrone motif several were selected for experimental validation. The selected targets should be pharmaceutically relevant, belong to different protein families, and biochemical and/or biological assays should be available. Moreover, incorporation of at least two out of the five scaffolds represented in the γ-pyrone branch shown in Figure 23 in different known inhibitor structures was required. Guided by these principles six proteins were selected, namely monoamine oxidases A and B, long investigated targets of anti-depressants[94-96], the signal transducers and activators of transcription (STATs) 1, 3 and 5B proteins that are promising drug targets in cancer[94-97 and inflammation[97], and the acid sphingomyelinase, a protein involved in apoptosis[98], Niemann Pick A/B disease[99] and secondary infections in cystic fibrosis. Initial pre-screens of the γ-pyrone library at fixed compound concentration identified potential hits that were validated in subsequent concentration-dependent measurements (see experimental sub-sections 1.6.4 - 1.6.6 for details). The results of the pre-screens are displayed as a heatmap in Figure 24c with percent of remaining enzyme activity ranging from 100% (black, no inhibition) to 0% (green, full inhibition). The heatmap identifies several potential inhibitors for the different target proteins possessing remarkable selectivity.

The monoamine oxidases (MAOs) subtype A and B catalyze the oxidative deamination of a range of monoamines including the neurotransmitters histamine, dopamine, noradrenalin and adrenaline. The isoforms differ in their sensitivity towards different inhibitors as well as in their substrates. MAO A oxidizes 5-hydroxytryptamine (serotonin) while MAO B converts benzylamine and 2-phenyethylamine. Both enzymes are targets in anti-depressive drug development and in the therapy of Alzheimer's and Parkinson's diseases.[94-96] Inhibitors for MAO A and B with varying selectivity have been and are still being actively developed.[100] The compounds from WOMBAT used for annotation of the γ-pyrone branch with inhibitory activity against monoamine oxidase are shown in Attachment 4.

The monoamine oxidase assays were established and carried out in collaboration with Dipl.-Chem. W. Wilk. Concentration-depenedent measurements of MAO A and B inhibitory activity using a fluorescence-based assays developed by Novaroli *et al*.[101] identified more than 60 and 35 hits for MAO A and B, respectively, with IC_{50} values below or equal to 10 µM. Of these inhibitors 25% (15 compounds) and 31% (11 compounds) exhibit IC_{50} values below 1 µM for MAO A and B, respectively. Many of the identified inhibitors showed selectivity for the corresponding isoenzyme to some degree. Seven inhibitors showed more than a hundred fold selectivity for MAO A over MAO B whereas six showed inverse selectivity in the same range. A selection of hit compounds is shown in Table 7, for all hits see Attachment 5.

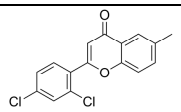
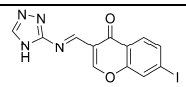
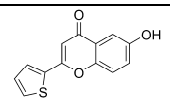
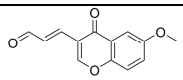
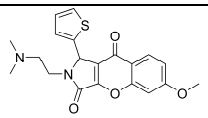
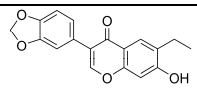
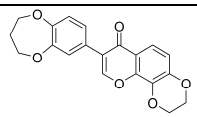
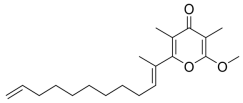
Cpd.	Structures	IC_{50} [µM]		LBE
		MAO A	MAO B	
GPL-300		0.95 ± 0.07	> 50	0.30
GPL-18		1.31 ± 0.04	> 50	0.31
GPL-4		0.94 ± 0.03	> 50	0.35
GPL-76		3.96 ± 0.19	> 50	0.32
GPL-75		> 50	0.34 ± 0.17	0.28
GPL-397		> 50	0.31 ± 0.03	0.24
GPL-351		> 50	0.54 ± 0.06	0.24
GPL-458		> 50	4.88 ± 0.18	0.23

Table 7: Selected hits from the monoamine oxidase screen with measured IC_{50} values and ligand-binding efficiency. The compound identifiers refer to those used in Attachment 3.

A closer analysis of the γ-pyrone structures annotated as active against MAO A or B in WOMBAT that were used for the target annotation revealed mainly compounds embodying an isoflavonoid scaffold as active compounds but much fewer flavones or xanthones (see Attachment 4). Interestingly, the screen identified compounds from all scaffold families as hits whereas particularly high hit rates are observed for the xanthones for both monoamine oxidases and for the flavones for MAO A. In the light of these findings, both scaffold classes emerge as promising starting points for the development of monoamine oxidase inhibitors. However, the average ligand binding efficiency (LBE)[85] of the xanthones is low, most likely because of the large molecular scaffold. From this perspective, the chromone scaffold family may represent a more promising starting point with higher LBEs and a hit rate of 5 % against both monoamine oxidases. Compared to an average hit rate of 0.34% over all assays stored in PubChem and taking into account the generally higher hit rates of focussed libraries, the hit rates observed in the monoamine oxidase screens can be considered as very high. With tens of active compounds, a structure-activity relationship (SAR) should be established for the γ-pyrones.

However, as laid out before, substituent diversity was one key criterion guiding assembly of the library which greatly complicated the delineation of SAR patterns. Therefore, a SAR could only be established for the xanthone scaffold for which substitution at positions two and four reduced inhibitory potency for both monoamine oxidase whereas substitution at position three only drecreased activity against MAO A. (see Figure 25).

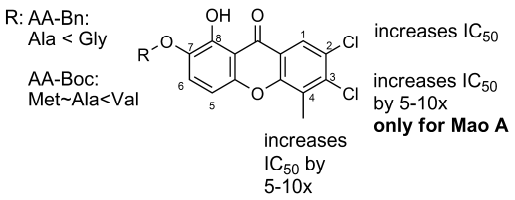


Figure 25: SAR for the xanthone scaffold-derived hits in the monoamine oxidase screen.

As second target protein group the signal transducers and activators of transcription (STAT) were chosen. STATs are latent cytoplasmic transcription factors that bind to activated cytokine or growth factor receptors via a conserved Src homology (SH) 2 domain. This binding event is prerequisite to the subsequent phosphorylation of STATs at a conserved tyrosine by receptor-associated tyrosine kinases or the intrinsic kinase activity of growth factor receptors. Subsequently, tyrosine-phosphorylated STATs form a particular kind of dimers by reciprocal *p*Tyr-SH2 interactions, and accumulate in the nucleus, where they regulate expression of their target genes. Constitutive or upregulated activity of phosphorylated STAT1, STAT3 and STAT5 has been linked to several cancers including breast and prostate cancer as well as leukemia.[102] In particular, STAT3 and 5a/b are being actively investigated as anti-cancer drug targets because they are constitutively tyrosine-phosphorylated (and thereby activated) in a large proportion of human cancer cells[103,104] and inhibition of signaling via STAT3/5 in these cells has been uniformly shown to induce apoptosis.[105,106] STAT3 has also been identified as an important factor for angiogenesis, i.e., the formation of new blood vessels, a decisive process in tumor growth.[107-109] The STAT1 pathway is also a promising candidate for the development of novel anti-inflammatory drugs possibly lacking the side-effects associated with chronic use of non-steroidal anti-inflammatory drugs.[97] Known small molecule inhibitors mainly target STAT3,[106] and STAT5. Chromone-based inhibitors binding to STAT1 have been described in literature. This activity was used to annotate the γ-pyrone branch (see Attachment 6 for structures and activities).[110] The screen (see the Experimental section) revealed compounds **REF**, **GPL-5b** and **GPL-74** as selective STAT inhibitors (see Table 8). The STAT assay and the western blot experiments were conducted by B. Sperl and Dr. T. Berg at the Max Planck Institute of Biochemistry in Munich.

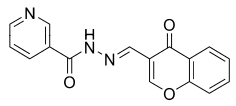
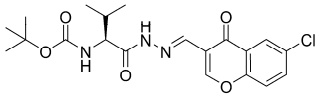
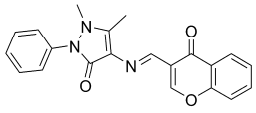
Cpd	Structure	app. IC_{50} [µM]		
		STAT5b	STAT3	STAT1
REF		32 ± 1	70 ± 3	> 500
GPL-5		38 ± 2	30.5 ± 0.4	29 ± 1
GPL-74		14 ± 1	22 ± 1	39 ± 1

Table 8: In vitro activities of γ-pyrone hits against STAT proteins determined by fluorescence polarization assays. The known inhibitor **REF**[110] was included in the screen to ensure comparability of the results. The hits discovered in this screen posess more promising properties (LBE and AlogP) than those from WOMBAT. All compounds of the γ-pyrone library (GPL) were numbered sequentially as shown in Attachment 3. The reference inhibitor was labeled "REF" to differentiate it from the library.

Cellular experiments were performed to confirm the biochemical assay results for the compounds **REF** and **GPL-5** (see Figure 26). The imine moiety of the Schiff base **GPL-74** is probably hydrolyzed in the buffer or within the cells rendering this compound inactive at high concentrations. Compound **GPL-5** was experimentally determined to be stable for at least three hours under assay conditions, i.e., well within the time required for the experiment. Compounds **GPL-11** and **GPL-74** are similar to a set of STAT inhibitors recently discovered by Berg *et al*. in a high throughput screen[110], albeit with different selectivity patterns (see Attachment 6). The compounds identified by the screen presented here and the screen of Berg *et al*. are among the first small molecule inhibitor classes of STAT1.

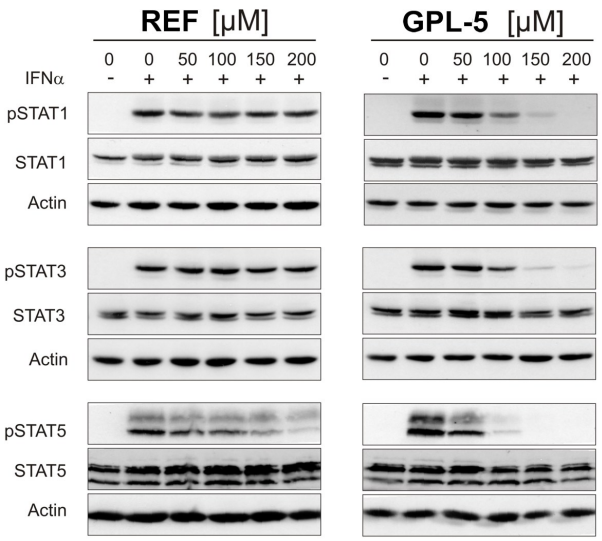


Figure 26: Cellular activities of the inhibitors REF and GPL-5 as determined by Western blotting with *p*Tyr specific antibodies. Since STAT phosphorylation requires prior STAT binding to activated receptors by the STAT SH2 domains, an inhibitor of STAT will lead to reduced levels of tyrosine-phosphorylated STAT protein.

The γ-pyrones were also screened against acid sphingomyelinase. The sphingomyelinases catalyze the cleavage of sphingomyelin to phosphorylcholine and ceramide, a second

messenger involved, among others, in regulation of differentiation, proliferation and apoptosis.[98] Neutral sphingomyelinase has been linked to TNF_{α}-mediated apoptosis and more recently to the formation of exosomes[111], which are thought to play a key role in cell-cell contacts and the pathogenesis of retroviral infections.[112] A partial deficiency of acid sphingomyelinase has been discovered to cause Niemann Pick disease Type A and B, a hereditary metabolic disorder that leads to accumulation of lipid metabolites.[99]

The acid sphingomyelinase recently emerged as a potential drug target, since it holds a key role in the development of platelet aggregating factor (PAF) induced lung edema leading to acute lung injury (ALI, a major cause of death for intensive care unit patients),[113] and lung emphysema.[114] In addition, it has been reported that inhibiting acid sphingomyelinase prevented susceptibility to bacterial infections in a mouse-model of cystic fibrosis.[115] In contrast to the neutral sphingomyelinase, for which some potent inhibitors are available[116-119] there is only one potent and selective inhibitor known for the acid sphingomyelinase.[120] The latter however, is not cell permeable and thus only active in cell-free assays. For acid sphingomyelinase a fluorescence-based assay was used and inhibition of neutral sphingomyelinase was determined by a radioactivity-based assay. Both assays were run by Dipl.-Chem. A. Roth and Prof. Dr. C. Arenz at the Humboldt University in Berlin in collaboration with A. Yektaoglu (M.Sc.) and Prof. Dr. A. Giannis at the University of Leipzig.

The screen revealed two potent compounds with a benzopyran scaffold (see Table 9, compounds **GPL-229**, **GPL-61** and **GPL-214**). The IC_{50} values of the best inhibitor (**GPL-229**) was determined to 3.1 µM for acid sphingomyelinase with a selectivity of more than 15-fold over the neutral isoenzyme. Although the only known small molecule inhibitors of sphingomyelinase including the natural product α-Mangostin (**WB-1**)[121] all embody the xanthone scaffold and guided the bioactivity annotation of the γ-pyrone branch (see Table 9, compounds **WB-1** to **WB-3**), this screen did not yield hits from this scaffold class. From the available data, it was hypothesized that the prenyl moieties shift selectivity towards the inhibition of acid sphingomyelinase.[122] The prenylated compounds are highly lipophilic whereas the selective inhibitors newly discovered by the screen posess better potency and much lower lipophilicity, thus presenting better starting points for further optimization.

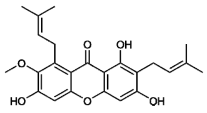
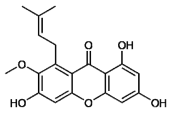
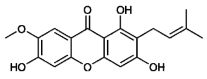
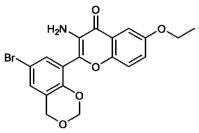
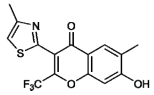
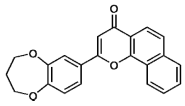
Number	Structures	IC_{50} / K_i [µM]		LBE	AlogP	Ref.
		acid	neutral			
WB-1		12.5	113.3	0.16	5.9	[122]
WB-2		14.6	24.7	0.19	4.1	[122]
WB-3		108	> 202	0.16	4.1	[122]
GPL-229		3.1 ± 2.1	>> 50	0.21	2.6	
GPL-61		9.5 ± 2.8	>> 50	0.22	3.4	
GPL-214		21.5 ± 6.8	> 50	0.18	3.9	

Table 9: Results of the screens for sphingomyelinase inhibition. Compounds **WB-1** to **WB-3** were taken from the WOMBAT database and used to identify sphingomyelinases as potential targets of the γ-pyrone branch. The remaining entries display the inhibitors identified by screening and their IC_{50} values. To determine the selectivity of the compounds over neutral sphingomyelinase, the corresponding IC_{50} values were determined as well. Ligand binding efficiency (LBE) and AlogP values are also given for all compounds. The compounds from Wombat (WB) and the γ-pyron library (GPL) are numbered according to the corresponding tables Table 9 and Attachment 3, respectively.

1.4 Discussion

1.4.1 SCONP 2.0: Rules for exploration of chemical space

The new set of rules developed together with A. Schuffenhauer and P. Ertl from Novartis[56] improved the usability of the scaffold tree concept by introducing “virtual scaffolds”, i.e., scaffolds that do represent compounds in the analyzed data set but are generated during the scaffold pruning procedure. There are three major benefits from this: first, virtual scaffolds guarantee the ‘structural integrity’ of the scaffold tree and prevent holes without any structure. Thus, the step-wise deconstruction of scaffolds is also reflected in the diagram rendering the scaffold tree more intuitive. Second, the missing gaps ensure independence of the data set when constructing a scaffold tree since the selection of the parent scaffold exclusively depends on the set of rules and not on the data set analyzed, i.e., the same molecule always results in the same scaffold branch. In particular, scaffold tree merging as applied in the biological annotation of the γ-pyrone branch of the natural product tree became possible. In the previous procedure where the resulting tree diagram depended on the data set analyzed this analysis would have been impossible. Comparison of sets of compounds is a self-evident application of any classification system of chemical space including scaffold trees. Third, the virtual scaffolds themselves represent opportunities since they define gaps in the analyzed chemical space that can be filled with molecules through synthesis or acquisition. Combined with ‘brachiation’, the propagation of molecular properties (e.g. bioactivity) along branches of the scaffold tree, virtual scaffolds may allow to identify novel inhibitor types, as shown in section 1.3.5. for the discovery of novel kinase inhibitors.

Although this set of rules was developed carefully and based on synthetic and medicinal chemistry knowledge and successfully applied in the development of novel inhibitor types, it is neither the “best” nor the “only” one. The definition of the “best” set of rules depends on the purpose of the generated scaffold tree, i.e., on the insights sought and the questions to which answers are needed. Especially rules incorporating knowledge and experience are highly subjective. Rule 6 prioritizing the removal of rings with size 3, 5 and 6 or rule 9 setting the heteroatom priority N > O > S (see Figure 8) are examples that may yield good results for analyses of medicinal chemistry compounds and for pharmaceutically relevant questions. However, they may not be suitable in analyses aiming at other purposes. Therefore, the set of rules applied in this work proposes one solution based on the rational of synthetic and organic chemistry. They do not exclude that there are other sets of rules that perform similarly well or even better, depending on the analysis at hand. The Scaffold Tree Generator offers the functionality to customize the rule set and, thereby, explore the possible alternatives for any given data set with reasonable effort. A more dynamic approach would allow changes in the rule

set ‘online’ and following the changes in the scaffold tree in real time. Both approaches and their implementation are discussed in sub-section 1.5.2.

1.4.2 Scaffold Hunter – a versatile tool for chemical space exploration

Scaffold Hunter introduces a new concept for chemists and biologists to interact with their data by directly linking numbers to chemical structures and to dynamically visualize correlations. Thus it provides an intuitive and rapid way to assess and analyze large amounts of data, which will enable scientists who are experts in their field to access their data more rapidly. Its intuitive and straight forward analysis capabilities prevent the analysis being a ‘black-box’ to those scientists and researchers depending on and working with the results. This makes Scaffold Hunter a unique application for the evaluation of data in biomedical research.

Scaffold Hunter’s modular code and the generic data structure ensure applicability to a wide variety of data irrespective of their source or classification with minimal modifications. Thus Scaffold Hunter will improve the conversion of data to knowledge and, ultimately, to science, and new hypotheses. This will benefit many projects at the interface of chemistry and biology, e.g., in chemical biology, biochemistry or medical research. Scaffold Hunter has been published[123] and is available free of charge from http://www.scaffoldhunter.com under an open source license. Scaffold Tree Generator, the program for building the scaffold tree database, is available from the same source under the same conditions. Both programs have been implemented using free software exclusively rendering them free as well. The well-documented source code for both applications is available, so that Scaffold Hunter and Scaffold Tree Generator can be applied, maintained and extended by researches everywhere, in academia and industry, according to their needs. This offers practically unlimited flexibility to the users and, hopefully, will foster the use of Scaffold Hunter throughout academia and industry.

Some features of Scaffold Hunter may be found in very expensive and highly specialized software packages related to cheminformatics, high throughput screening evaluation and structure-activity-relationship (SAR) delineation. Nonetheless, the combination of an instructive visualization, especially with regards to chemical structures, and data mining functionality is not readily available in most commercial software packages.

Although some basic training of a few hours helps the users to become accustomed to the program, its use is quite intuitive and, in principle, it can be applied by every researcher. This is an important point since research is a creative process that strongly depends on the feedback loop between the scientist and his experimental data. Thus, data analysis should remain a task of the researcher who often is an expert in his own field rather than in data mining. In many research environments, e.g. academia or small and medium sized businesses, the researcher has no choice but to evaluate his own data simply because data mining experts are not available. Here, Scaffold Hunter could make a decisive impact by easing at least the initial

stages of data analysis and thus shorten the time for the extraction of knowledge from data while possibly improving the understanding of general trends as well.

1.4.3 Filling the gaps: discovery of novel pyruvate kinase inhibitors

For the efficient analysis of information-rich datasets obtained, e. g., from biochemical or biological screening of large compound libraries and the navigation through the associated chemical and biological space, powerful cheminformatics methods are in high demand that foster recognition of structural relationships associated with bioactivity. Ideally such methods would be highly interactive and intuitively accessible to the educated non-specialist (i. e. to chemists and biologists with non-expert knowledge in cheminformatics). They should facilitate easy and interactive navigation through large data sets and the chemical space associated with them. Such methods would be of particular value if they would also allow the identification of novel scaffold and compound classes endowed with the desired activity. The results detailed in section 1.3.5 demonstrate that the Scaffold Hunter fulfils these criteria.

Scaffold Hunter has been developed with these two goals in mind: efficiency – in order to rapidly navigate through large chemical and biological spaces, and ease of interpretation – in order to offer an approach that is both intuitive and chemically relevant. The program automatically reads chemical and bioactivity data, generates library scaffolds, annotates them with bioactivity and arranges them in a hierarchical scaffold tree according to a set of rules derived from chemistry and medicinal chemistry criteria. The Scaffold Hunter is easy to use, highly interactive and facilitates direct, user-oriented intuitive recognition of structural relationships encoded with bioactivity data. Through brachiation along the branches of the scaffold tree, i. e. along lines of biological relevance, it allows the identification of virtual scaffolds, which represent gaps in the chemical space covered by the analyzed compound library.

Analysis of PubChem data and cross-correlation with the WOMBAT database convincingly demonstrate that such gaps are valid starting points in chemical space for the development of novel scaffold types endowed with activity for the protein target investigated. Scaffold Hunter analyses can also be employed to guide prospective compound development as demonstrated by the brachiation-guided identification of novel scaffolds for pyruvate kinase inhibitors and activators. The hit rate of 10% is extraordinarily high compared to average hit rates in PubChem HTS campaigns that on average yield 0.34% hits. The hit rate of the PubChem screen employed by this analysis was 1% at a 10 µM cut-off which is still ten-fold lower than ours.

It should be noted that Scaffold Hunter is not designed for the delineation of classical structure-activity relationships (SAR), which are often dictated by the nature and properties of substituents attached to the scaffolds of active compounds classes. These are pruned in the construction of the scaffold tree.

Scaffold Hunter's strength is the identification of promising structural templated for the deisgn of focused compound libraries, that is, the generation of hypotheses guiding future synthesis and, in general, research efforts. The structures identified from promising virtual scaffolds that represent gaps in chemical space are templates but usually do not constitute compounds or inhibitors. A compound collection is generated around the given template by addition of a diverse set of substituents. Since in most cases the substituents interact with the protein whereas the scaffold primarily orients the substituents in space, a diverse set of substituents is crucial to match the diversity in the interacting residues of the protein targets.

The pyruvate kinase inhibitor example provides a proof-of-concept albeit general applicability cannot be claimed. Although compound collections based on virtual scaffolds will possibly not yield inhibitors in each and every case, they facilitate the successful discovery of novel small molecule modulators of protein function for a fraction of targets.

Naturally, one would assume that inactivity propagates within a given scaffold branch analogous to activity. It is important to note, however, that brachiation based on active compounds will exclusively identify branches and scaffolds enriched with biochemical activity. Other branches may contain scaffolds representing active compounds as well since typically multiple chemotypes can bind to the binding sites of a given protein.

Besides the identification of virtual scaffolds to guide compound library development, Scaffold Hunter may prove a powerful tool in library management. The program facilitates the mapping of compound collections onto one another and the quick identification of structural overlaps guiding future library extension by either synthesis or compound purchase. Scaffold Hunter is also used as a chemically intuitive and quick means to prioritize hit lists of high throughput screening campaigns and select promising compounds for subsequent studies.

1.4.4 Annotation of biochemical activity by scaffold tree merging

Natural products form a compound class endowed with special properties and biologically prevalidated by their ability to bind to various proteins in their natural environment. Moreover, many of the natural products exhibit drug-like properties, for instance the majority of natural products passes the Rule-of-Five,[3,10,12,124] whereas some natural products can even be considered lead-like. Thus, natural products form an optimal starting point in chemical space for the development of bioactive small molecules as biological tools and drugs, e.g. by biology oriented synthesis (BIOS).[44] Nonetheless, their value is greatly reduced by the lack of detailed information about the individual bioactivity and the corresponding protein targets. Methods for the prospective assignment of potential targets are activily investigated.

The scaffold tree approach has proven its value as a means to chart and navigate through chemical space, as well as to guide chemical synthesis by brachiation and the identification of virtual scaffolds enriched with bioactivity for the protein target of interest.[23,56,125,126] Even though scaffold trees in combination with Scaffold Hunter are a valuable tool to guide chemistry and

biology research, the identification of promising compounds for specific protein targets still relies on activity information that is mostly not available for natural products.

Scaffold Hunter had been applied before to compare different compound libraries by co-mapping them onto one merged scaffold tree. Similarly, the annotation of biological activity via co-mapping of collections of non-natural compounds annotated with bioactivity and target information could provide the missing link between natural product scaffold structures and potential protein targets.

This concept was examined by generation of a merged scaffold tree from the Dictionary of Natural Products (DNP) and the WOMBAT database that lists literature bioactivity and target data for about 190,000 mostly synthetic molecules. This 'gigatree' comprises of 137,000 scaffolds populating 4,600 branches. An analysis for already validated brachiation sequences, i.e. series of scaffolds in a branch that represent compounds with experimentally proven activity against the same protein target, yielded numerous examples up to the length of four. In line with the findings of Renner *et al*.[89,92] this illustrates the broad applicability of the brachiation concept. Beyond that, the analysis also proved for the first time the good performance of the exclusively chemistry-guided set of rules with respect to brachiation.

For prospective experimental validation of the target annotation the branch of the γ-pyrones was selected, a natural product family represented well in both, the DNP and WOMBAT. By merging both scaffold trees, the γ-pyrone branch from the DNP was annotated with bioactivity and target information from the WOMBAT compounds. This step immediately led to the identification of eight brachiation sequences already present in the WOMBAT data that involved compounds from at least three different scaffolds in the branch indicating that the target annotation may as well extend from the annotated scaffold to the neighbours in the branch, thereby identifiying potential templates for the design of natural product-derived compound libraries targeting one particular protein.

To evaluate this concept, a diverse 500-membered γ-pyrone library was assembled including compounds from different hierarchy levels and branches of the scaffold tree. From the target annotation the monoamine oxidases A and B were selected, the STATs 1, 3 and 5b, as well as sphingomyelinase as promising examples to screen the library against. This screen yielded submicromolar selective hits for the monamine oxidases at a very high hit rate. Furthermore, two STAT inhibitors were discovered, one of which also proved active in a subsequent cell-based assay. The sphingomyelinase screen identified the first monobenzopyrane-based inhibitors but no xanthone-containing compounds which were the structural motif guiding the target annotation. Most hit compounds exhibit molecular properties well within the accepted ranges, e.g. the Rule-of-Five and lead-likeness criteria. Of particular interest is the pronounced selectivity that was observed in the hit compounds: each hit targets only one of the three protein families tested and in many cases even only one isoenzyme from this family.

Brachiation was observed in the monoamine oxidase screen since the discovered hits span almost all scaffold families tested. This may be an extreme case and brachiation was not observed in all cases, for example for the STATs or sphingomyelinase. Nonetheless, the results for monamine oxidase and other, previously reported examples[44,87,127], as well as the analysis of the WOMBAT database prove brachiation a valid concept in the search for structurally simplified natural product-based inhibitors.

Moreover, scaffold trees were successfully used to annotate protein targets to various, unannotated compounds via scaffold tree merging. It should be noted, however, that the target annotation is passed on via the scaffold not via individual compounds. This introduces some fuzzyness to the annotation since scaffolds orient substituents in space whereas most of the molecular interactions with the protein are formed by the side chains. Several consequences arise from these observations. First, once a scaffold has been identified that is linked to the target of interest substituent diversity of the assembled library is a key requirement for the successful discovery of potent compounds. This due to the fact, that only sufficient diversity grants a chance for an optimal match between the substituents that are preorganized in space by the scaffold and their interaction partners in the protein binding site. A minimum library size of 100 – 200 compounds is required for the same reason. Second, as shown in this work, selective compounds can be obtained from a diverse library guided by propective target annotation via scaffold trees. Although the pre-validated scaffold enhances the probability of a good match between the compound and the protein binding site leading to tighter binding, that is, it enriches the library with potential activity against the target of interest, the substituents form most of the interactions. At atomic resolution, the molecular interactions of protein binding sites do differ from each other, thereby selecting those compounds whose side chains offer the best match. Third, the number of domain architectures developed during evolution is quite limited in comparison to the total number of proteins.[43,128,129] The target annotation via scaffold trees is based on the complementarity between the molecular scaffold orienting the substituents in space and the corresponding sub-fold of the protein binding site which preorganizes the interaction amino acid side chains. Thus, the limited number of sub-folds indicates that prospective target annotation via molecular scaffolds is feasible. Moreover, it could imply that certain scaffolds target particular binding site types extremely well – a notion closely related to the concept of ‘priviledged structures’. Target annotation via scaffold trees may then be well suited to identify these ‘priviledged scaffolds’ as well as their corresponding target proteins. Fourth, interesting target clusters for a given library may be identified from analyses of structural similarity between binding sites (see Protein Structure Similarity Clustering, section 2).

1.5 Outlook

1.5.1 Scaffold Tree Generator: from chemistry-based to biology-derived scaffold trees

The chemistry-bases rules set as described in section 1.3.2 has been implemented in Scaffold Tree Generator and, thereby, became available to the broad scientific community for free under an open source license. It comprises the initial set of rules developed in collaboration with A. Schuffenhauer and P. Ertl from Novartis[56] and is used throughout this work. In addition to this set of rules, Scaffold Tree Generator offers the possibility to create and apply user-defined rule sets where the rules and their order can be freely chosen without any changes in the source code of the program itself. The rules are limited to the 40 molecular properties currently defined inside Scaffold Tree Generator. More criteria can be added by extending the source code of the program – not as easy as using the implemented ones but with a reasonable effort. This mechanism offers a way to adapt the set of rules to the question or the data set at hand. Enabling modification of the set of rules through a graphical user interface would further improve the applicability and can be envisioned for future versions of Scaffold Tree Generator since the implementation should require only moderate effort and time. With a graphical representation of the rule set, modification should be possible for the educated non-expert, i.e., chemists and biologists, enabling them to create scaffold trees according to their needs. Ideally, a 'barcode' would automatically be generated for each rule set that describes the rules used and their order. Such 'barcodes' could easily be transmitted via e-mail and would enable other users to employ the same set of rules with a minimum effort.

Additionally, there may be other methods for construction of scaffold trees and, hence, to charting chemical space than chemistry-based methods. In close collaboration with the Scaffold Hunter project, Steffen Renner developed a biology-based method for the construction of scaffold trees.[89] In brief, Renner *et al.* generated all possible parent scaffolds from a given scaffold using the ring pruning procedure developed before.[56] Parent-child-scaffold pairs are formed from the child scaffold and those parent scaffolds that represent molecules in the data set (See Figure 27). Once all parent-child pairs have been generated, branches can be built in iterative steps from them by identifying and combining pairs where the child scaffold of one pair is the parent scaffold of the other one or *vice versa* (see Figure 28). In a final step, only those sequences are retained that are annotated with biochemical or biological activity against the same molecular target. The result of such an analysis can be further analyzed at several levels of abstraction: target, target class/family and data set. Viewing the sequences generated for a given target in most cases leads to a scaffold tree comprising of a few branches with active molecules. Naturally, when extending the view to a target family, the tree will grow bigger, possibly including branches with multiple activity annotation, i.e., sequences inhibiting several (closely) related targets of the family. The view on a whole compound library incorporating several target classes may shed light on the promiscuity of several structure families and

identify privileged structures conferring modulatory activity against several proteins – one step towards directed poly-pharmacology and, finally, multi-targeted therapeutics.

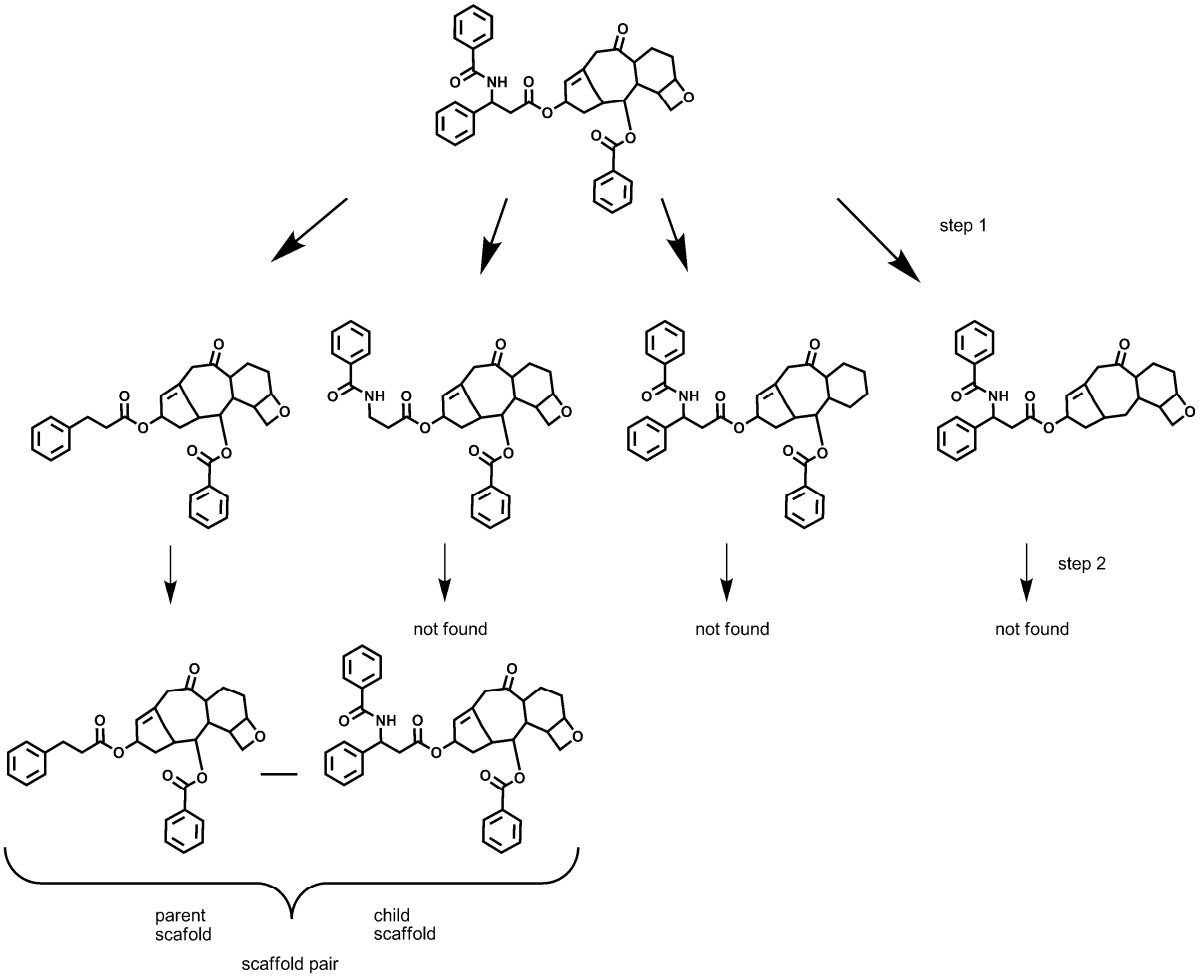


Figure 27: Identification of parent child scaffold pairs in the dataset. For each scaffold all possible potential parent scaffolds are generated, as defined by the ring pruning procedure used by Schuffenhauer *et al*.[56] Only the parents that are found as scaffolds from molecules in the original dataset are retained. Figure reproduced from Renner *et al*.[89]

The integration of the tree-building method developed by Renner *et al.* and described above may be a worthwhile endeavour since it complements the exclusively chemistry-based scaffold tree generation. As more and more chemical data sets annotated with high quality structure-related biology information become available, e.g. PubChem[130] or StARLite[131], biology-based

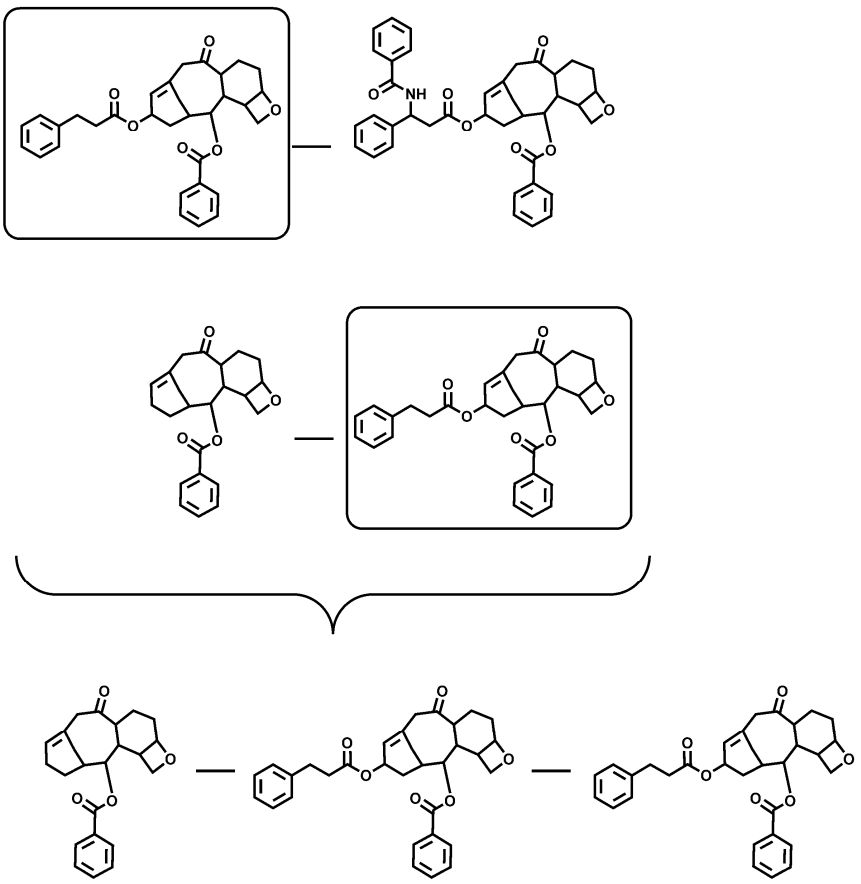


Figure 28: Generation of linear branches from the parent child scaffold pairs. Two branches (e.g. pairs, triplets, …), where the smallest scaffold in the one branch is the same as the largest scaffold in the other branch, can be merged into a longer branch containing all the scaffolds from the smaller branches. Figure reproduced from Renner *et al*.[89]

scaffold trees could provide one intuitive way to use this information in the quest for novel small molecule modulators of protein function. The efforts needed to extend Scaffold Tree Generator in this direction are estimated to require several weeks of work by a specialist. The initial modules for parent scaffold generation already exist within the Scaffold Tree Generator code. A module filtering for those scaffolds representing molecules in the data set would need to be implemented. Subsequent branch building and annotation with biological activity would need to be implemented as well.

To effectively use such trees, the implementation of string property-based, i.e., target names or UniProt[75] IDs, in Scaffold Hunter would be needed. So far, filtering is exlusively based on numerical properties.

1.5.2 Possible extensions of Scaffold Hunter and its scope of application

It has been shown that Scaffold Hunter provides an intuitive classification of chemical space together with a user-friendly graphical interface that enables non-expert users to look at and analyze their data. Nonetheless, there is still a lot of room for improvement since different

analysis tasks require different functionalities. As pointed out in sub-section 1.4.2, the multitude of possible and useful features is almost unlimited and this was one of the reasons to make Scaffold Hunter available as open source: to integrate the community into the development of the program, thereby, possibly speeding up and diversifying development. From extensive testing internally and in pharmaceutical companies as well as presentation at scientific meetings several ideas arose, how Scaffold Hunter could be improved and extended. These ideas, as well as potential ways to implement them, will be discussed in this sub-section.

Finding the “right” scaffolds in Scaffold Hunter relies almost exclusively on the filtering according to numerical properties averaged over molecules represented by each scaffold and attributed to the corresponding scaffold. Although it was shown that these averaged properties can indeed be successful for the discovery of novel, potent inhibitor families, in other cases filtering by molecular properties may be more interesting. This may be the case when the filter is applied to physico-chemical properties, for example, that depend largely on the substituents. One may argue that this is also true for biochemical and biological activity. However, in this case, the information about inactive molecules represented by the same scaffold is very valuable and should not be ignored, whereas for physico-chemical properties, the molecules failing the filter can be omitted. Implementing a possibility to filter on the basis of molecular properties and display the scaffold tree built from all the molecules passing the filter is certainly doable. First of all, this would require a module to filter the molecular properties resulting in a list of structure ids. The module for filtering according to scaffold properties could be adapted to this purpose and return a list of structure ids. An additional database query for the scaffold ids of these compounds would generate a list of scaffolds ids with which a tree could be built.

The proposed feature of filtering according to structure properties also touches upon the identification of structures by their canonical SMILES string. Although that directly relates to chemistry, it also generates problems – especially database searches are slowed down since SMILES strings can get quite long and indexes are often built on only the first 20 characters of a string. Moreover, canonical SMILES differ depending on the program used to generate them which may lead to problems with the uniqueness of the identifier. Therefore, it may be advantageous to replace the strings as unique identifiers by a numerical structure id by analogy to the scaffold id. Although such a step would require changes in the database structure, the Scaffold Tree Generator program and Scaffold Hunter, the effort seems reasonable compared to the potential benefits.

Another way to extend Scaffold Hunter’s functionality that was often requested by users is the addition of a substructure search engine. So far, Scaffold Hunter is devoid of any chemical intelligence, e.g., software modules that relate to chemical structures or chemical properties. This was done on purpose since Scaffold Hunter should be kept as generic as possible and all the needed chemical intelligence was, therefore, implemented in the Scaffold Tree Generator.

However, a substructure search would significantly increase the usability for chemical space exploration. One could image two possible application scenarios for substructure searching: in the scaffolds or in the underlying compound structures themselves.

Searching for a given structural motif in the scaffolds obviously would help to find scaffold classes of interest to the user. It can also identify those scaffolds present in the analyzed data set that embody certain structural motifs, for example known pharmacophoric substructures mediating binding to the protein of interest. The search would ideally select all the found scaffolds and thus allow to use the “selected scaffolds in new tab” function and to directly arrive at a selected sub-tree.

Searching in the molecular structures underlying the scaffold tree would be of interest since such searches could incorporate aliphatic substituents that may be known as integral parts of a pharmacophoric motif mediating biological activity against the target of interest. Substructure searching within the molecular structures may, therefore, enhance Scaffold Hunters ability to identify structure-activity-relationships (SAR). Development of a substructure search engine based on open source technology is currently pursued by Nils Kriege in his master thesis. This search engine will feature a molecular editor for query input as well as wildcard searching, which will then be incorporated into Scaffold Hunter.

The set of rules used for selecting the single parent scaffold and, hence, building a tree-like data structure is subject to much discussion. Without this set of rules, the tree would change to a graph, i.e., the chemically meaningful scaffold of a molecule would be connected to all its parent scaffolds that, in turn, would be connected to all their child scaffolds. The resulting diagram would contain significantly more connections and readability would decrease to a minimum. Nonetheless, it could be interesting to visualize alternative routes of dissection, i.e., alternative parent scaffolds, in the scaffold tree, for example when one would like to optimize the set of rules or study the impact of changes in the rules and their order. Although such a feature may be useful for certain tasks, the implementation would require massive changes in Scaffold Hunter and its database structure. This would produce a highly specialized version that would not be downwardly compatible, lose its generic character, and evolve into a “chemical space graph visualization tool”. If there is enough demand for such a specialized version, it may be possible for a group of users to take the Scaffold Hunter source code and modify it accordingly.

Similarly, users have proposed a more flexible data structure that would not visualize a pre-computed tree but rather allow on-the-fly changes in the tree structure; manual assignments of child scaffolds to parent scaffolds as well as changes in the set of rules, i.e., removal or addition of rules or changes in their order. Such a feature would obviously allow highly dynamic analyses and adaptation of the set of rules guiding the tree generation. As for the feature laid out before, a massive change in data structure would be needed, combined with the reprogramming of

major parts of the Scaffold Hunter source code. The speed of such an application would be one of the most critical issues since memory usage is restricted and the scaffold tree needs to be rebuilt after every change within an acceptable time-frame. One possible solution that requires even more effort for implementation would be the re-design as client-sever-based application. Database operations as well as real-time tree building would be done on a server with more memory and computational power than a normal desktop computer. The pre-processed data would then be transferred for visualization to the Scaffold Hunter via a network connection and be displayed there. Such a concept could also utilize Scaffold Hunter with some minor extensions for communication with the application server module as the visualisation program. All the dynamic computation would be done by the newly developed server component. Design and implementation would definitely be a larger project but the overall potential and benefit-cost analysis seems promising.

So far, Scaffold Hunter has been applied to the identification of parts of chemical space enriched with biological relevance and the discovery of bioactive scaffolds and compounds. However, other compound classes can be classified as well like, for instance, catalysts or odorous compounds. Possibly, virtual scaffolds in such scaffold trees can be explored to discover novel catalysts or scents, respectively. Although Scaffold Hunter was planned and implemented as a visualization tool for the exploratory navigation through various chemical spaces, the underlying concept is also suitable for similar problems from other domains, e.g., in bioinformatics. The modular and generic structure enables the application to data from other domains, e.g., protein structure classifications or functional hierarchies, with minimal modifications. The structures that Scaffold Hunter displays on the nodes are stored as an SVG image and, therefore, any other graphical representation that can be stored as an SVG image may serve as graphical template. Modifying Scaffold Hunter for the visualization of protein structure hierarchies as, for example, defined in the Structural Classification of Proteins (SCOP) database[132] would satisfy all the necessary criteria (tree-like data structure, multiple branches) needed for successful visualization in Scaffold Hunter. One issue remains: the generation of protein structure SVG images for which no program could be discovered during extensive web searches. Extension to hierarchically classified data from other sources should be possible as well.

On the application side, the successful examples outlined in subsections 1.3.5 and 1.3.6 do not suffice to assume general applicability but give a fairly good impression of potential application scenarios of Scaffold Hunter in the quest for novel small molecule modulators of protein function. Public availability free of charge should encourage scientists in academia and industry to use Scaffold Hunter in their daily research. Application examples with early adopters may fuel further interest in Scaffold Hunter. To this end, an analysis of public databases like PubChem or StARLite could on the one hand add an interesting perspective on their data and, on the other

hand, integrate application and development of Scaffold Hunter into a larger movement towards publicly available data and methodology to foster biomedical research. Moreover, such analyses would yield further starting points in chemical space for the design of small molecule modulators of protein function for a plethora of targets, thereby bridging the gap between cheminformatics and applied chemical and biochemical research.

1.5.3 From scaffolds to fragments

Fragment-based drug discovery (FBDD) has developed over the past years as a new paradigm in drug discovery.[71,72,133-135] Its progress has been driven by two notions: first the insight that the molecular diversity covered by screening libraries of several million compounds is still relatively small compared to the chemical space of all putative medicinal chemistry compounds.[136] The second notion is that the binding probability of a molecule decreases with its size since the likelihood of a good match with the protein binding pocket decreases with increasing complexity of the molecule.[137]

Screening low molecular weight fragments, i.e., parts of larger molecules often derived from known bioactive molecules like drugs, addresses these notions. Due to their small size, fragments have a higher probability for a good match with the protein binding pocket. Moreover, by screening a number as small as several hundred or a thousand fragments, one addresses a theoretical chemical space comprising of up to a billion possible molecules comprising of three linked fragments each. In general, hits and leads resulting from FBDD tend to be more polar than hits from HTS and possess a much higher ligand binding efficiency.[138]

The scaffolds generated by the ring pruning procedure and used in the scaffold tree resemble fragments quite closely. This is reflected, for example, by the fact that 31% of the scaffolds in the natural product scaffold tree and 18% of the WOMBAT scaffolds pass the Rule-of-three filter defining empirically derived properties for fragment like molecules by analogy to the Rule-of-five. In brief, fragment-like molecules should have a molecular weight below 300 Dalton, three or less hydrogen bond donors and acceptors, and a calculated octanol-water-partition coefficient (ClogP) below three. Additionally, three or less rotatable bonds and a total polar surface area of 60 Å² or less may also be beneficial for fragment-like molecules.[139]

However, the scaffolds as occurring in the scaffold tree do not represent fragments according to the definition used here because they lack all side chains and functional groups therein that are important for forming molecular interactions with the protein and serve as anchor points for chemical modification in the fragment to lead optimization. One way to retain such functional groups could be the modification of the initial step where the chemically meaningful scaffold is isolated. An adapted algorithm can be envisioned that disassembles side chains in a retro-synthetic fashion analogous to the RECAP rules.[140] The scaffold with its attached functional groups would then be deconstructed in the same iterative fashion as in the scaffold tree. This procedure would be superior to the “normal” fragmentation of molecules into scaffolds, i.e., ring-

system generation or RECAP, since it constructs fragments of all sizes and also generates "virtual fragments" that may not have resulted from other methods. In conclusion, one could expect such a modified scaffold tree algorithm to yield an extensive scaffold library from a given set of molecules with a comprehensive coverage of the fragment space by virtual scaffolds filling potential gaps.

1.5.4 Exploiting nature's diversity: natural-product derived fragments

Today, fragment libraries have mostly been built on the fragmentation of drugs, leads and other compounds resulting from medicinal chemistry, thereby addressing those parts of chemical space that are assigned to drug space.[141] Although natural product chemical and property space partially overlaps with drug space, natural products occupy parts of chemical space that are mostly different from drug space. Natural product-based drug discovery has largely been abandoned throughout the pharmaceutical industry during the past decades, partly due to the notion that natural product are 'structurally complex' and 'difficult to produce'. [11]

Therefore, natural product-derived fragments may offer a promising means to explore the potential of natural products and address their chemical space. The synthetic effort should be reasonable since fragments are, in general, much smaller and less complex than molecules. Additionally, one would use a library in the size of several hundred up to maybe a few thousand fragments; far less compounds than needed for a successful HTS library. Most fragment libraries used in companies contain between one and two thousand compounds.[141]

Indeed, 31% of the natural product scaffolds pass the Rule-of-Three filter compared to only 18% of the WOMBAT scaffolds. This is a rather high fraction since, by analogy to the Rule-of-Five, the Rule-of-Three may not be unconditionally applicable to natural products because of their particular structural and molecular features, i.e., higher molecular weight as well as more oxygen than nitrogen atoms (resulting in more hydrogen bond acceptors and less donors).

Combined with the points raised in sub-section 1.5.3, natural products fragments could provide access to the diverse and biologically relevant natural product chemical space with a reasonable synthetic effort. Further fragment to lead optimization as well as lead optimization would then proceed as for the fragments derived from drugs or medicinal chemistry compounds. Since the natural product chemical space is highly complementary to the drug space, its exploration and exploitation may well lead to novel chemotypes for the modulation of protein function and, in the end, to novel drugs.

1.6 Experimental

1.6.1 Scaffold Tree Generator

The Scaffold Tree Generator software was initially written by Steffen Renner and later modified by Stefan Wetzel. In particular, errors were removed from the source code and its database interface was updated to be compatible with newer versions of Scaffold Hunter. Nils Krige contributed the graphical user interface for Scaffold Tree Generator. The program is available from www.scaffoldhunter.com. Scaffold Tree Generator was written in the programming language Java to ensure platform independence and generates a scaffold tree database from an SD File[4]. It requires the Java Runtime environment version 1.5 or higher, which can be obtained free of charge from http://www.java.com. The resulting scaffold tree data is automatically written either into an SD file or into a specified MySQL database (www.mysql.com) that can be queried by Scaffold Hunter. The scalable vector graphics (SVG) images of the molecules are also generated during the scaffold tree generation from the molecular coordinates stored in the input SD file.

By default, the implemented set of rules is the one published by Schuffenhauer *et al*.[56] However, the program also offers the possibility to modify and use a customized set of rules defined in a Scaffold Tree Generator specific format described in the documentation.

1.6.2 Scaffold Hunter: visualization of scaffold trees

The Scaffold Hunter software was implemented in Java using the Batik and Piccolo toolkits. It requires the Java Runtime environment version 1.5 or higher, which can be obtained free of charge from http://www.java.com.

Scaffold Hunter retrieves the scaffold tree data from a MySQL database and visualizes it. Instructions on installation and setup of the MySQL database, as well as a scheme describing the tables, indices and data formats of the database can be found at http://edoc.mpg.de/display.epl?mode=doc&id=429252. The database connection itself is established via a JDBC driver, a standardized database interface in Java. Together with the use

[4] The ctab format for storing molecular structures was initially developed by MDL Information Systems, now renamed to Symyx. The format uses text files and connection tables to describe molecular structure. The molecular description consists of two blocks: an atom block describing the atoms of the molecule together with their properties, e.g. charge, coordinates etc., and a bond block describing the bonds and their details, e.g., connected atoms, bond type. The definition is publicly available and can be downloaded here: http://www.symyx.com/downloads/public/ctfile/ctfile.pdf. While a mol file contains only one molecule, an SD file can contain multiple molecules together with various annotated properties. Due to the availability of the format definition and its early conception, SDF has become a standard file format for exchanging chemical structures.

of standard SQL for querying the database, this ensures a maximum of flexibility in the choice of the database system – in principle any SQL-based database for which a JDBC driver is available can be used with Scaffold Hunter, e.g. Oracle or DB2. However, the database tables and the indices for accelerated data retrieval have been optimized for MySQL and might need to be adapted for other database systems.

The Scaffold Hunter program itself consists of three modules dealing with database connection and data retrieval, visualization and tree layout, as well as user interface and interaction. The modular design facilitates easy orientation in the code fostering modification and extension.

The full Scaffold Hunter program code in its native and annotated form is available from http://sourceforge.net/projects/scaffoldhunter/. A full documentation in English is available at this site as well.

The databases used for the analysis of the pyruvate kinase screen in PubChem as described in sub-section 1.3.5 can be obtained from http://edoc.mpg.de/display.epl?mode=doc&id=429252.

The program was successfully tested under Windows XP and Vista, MacOS X and Linux. Depending on the memory size up to 1,500 scaffolds could be shown simultaneously. Initial retrieval of data from the database can sometimes be slower than usual and lead to waiting times.

1.6.3 Pyruvate kinase and lactate dehydrogenase assay

Lactate dehydrogenase was obtained from USB Corporation, Cleveland, USA. Pyruvate kinase from *Bacillus stearothermophilus*, ribose-5-phosphate (R5P), potassium dihydrogen phosphate, potassium chloride, magnesium sulphate, imidazole and phosphoenolpyruvic acid monopotassium salt (PEP) were bought from Sigma Aldrich. Adenosine-5'-diphosphate disodium salt (ADP), β-nicotinamide adenine dinucleotide (NADH) and DMSO were purchased from Serva Electrophoresis GmbH, Heidelberg, Germany. Sodium pyruvate was purchased from Alfa Aesar GmbH, Germany. All measurements were conducted in transparent 384 well small volume plates from Greiner with a total volume of 14 µL and carried out in triplicate using a Tecan infinite M200 plate reader set to absorbance at 340 nm. Plates were measured for 40 minutes. Pipetting was done using a Caliper Zymark Sciclone ALH 500 pipetting robot. All measurements were evaluated using Microsoft Excel and IDBS XLfit. The values reported are the average values and their standard deviation.

Pyruvate kinase assay: The lactate dehydrogenase (LDH) coupled pyruvate kinase assay was set up according to protocols from PubChem[76,77] and Sigma.[80] Firstly, 1 µL of compound solution in DMSO was dissolved in 49 µL of a solution containing all components except the pyruvate kinase substrate PEP. Of this mix, 4x7 µL were transferred to the small volume measurement plate. After incubation for 15 minutes at 30°C, 7 µL of PEP solution were added to each well and the plate was measured in a continuous kinetics mode. Final assay

concentrations were 50 mM imidazole (pH 7.2), 0.6 mM NADH, 0.4 mM ADP, 0.14 nM R5P, 50 mM potassium chloride, 7mM magnesium sulphate, 0.01% Tween 20, 0.2 U LDH, 0.07 U pyruvate kinase, 0.05% BSA and 2 mM PEP.

The data was treated as described by Inglese *et al*.[77] except for normalization. Activity data were normalized to a negative control without PEP instead of a negative control with the pyruvate kinase inhibitor luteolin.[77] Compounds with less than 50% or more than 130% activity in the pyruvate kinase assay were confirmed in a concentration dependent measurement using 11 concentrations with a 2-fold dilution starting at 100 µM. The IC_{50} values derived from these experiments are reported together with their standard deviation for three independent measurements.

Lactate dehydrogenase assay: To rule out false positives by inhibition of the LDH reporter system, all compounds were screened for possible inhibition of LDH with a protocol adapted from Sigma-Aldrich.[142] For this purpose, 1 µL of compound solution in DMSO was dissolved in 49 µL of a solution containing all components except pyruvate. Of this mix, 4x7 µL were transferred to the small volume measurement plate. After incubation for 15 minutes at 37°C, 7 µL of pyruvate solution were added to each well. Final concentrations in this assay were 38 mM potassium dihydrogen phosphate pH 7.6, 0.9 mM NADH, 0.01% Tween 20, 0.002 U LDH, 0.05% BSA and 1.2 mM sodium pyruvate.

1.6.4 Monoamine oxidase assay

Fluorescence Assay for MAO A and B established and carried out in collaboration with Dipl.-Chem. W. Wilk:[101]

Kynuramine, deprenyl, clorgyline, monoamine oxidase A and monoamine oxidase B were obtained from Sigma-Aldrich Chemie GmbH. Monoamine oxidase A and B (human, recombinant) were purchased as microsomes from baculovirus infected insect cells and stored at −78 °C. Plates were pipetted using single- and multi-channel pipettes from Eppendorf and a Caliper Zymark Sciclone ALH 500 pipetting robot.

For the initial pre screen, 1 µL of 10 mM compound solution in DMSO was dissolved in 49 µL of a solution containing all components except the substrate Kynuramine. Of this mix, 4x7 µL were transferred to the small volume measurement plate. After incubation for 10 minutes at room temperature, 7 µL of Kynuramine solution were added to each well. The increase of fluorescence was then measured at 30 °C for 30 minutes in a continuous mode. The final assay concentrations were 0.05 mM Kynuramine, 100 µM compound, 0.1% of NP-40 (detergent) and 0.1U/well MAO A or 0.125 U/well MAO B, respectively in a potassium phosphate buffer (0.1 M) at pH 7.4 made isotonic with KCl.

All compounds exhibiting less than 20% residual activity in the initial pre-screen at 100 µM compound concentration were tested in a concentration-dependent manner with 11

concentrations using 1:2 dilution starting from 100 µM. All inhibitors with IC_{50} values below 1.5 µM were re-measured in a concentration-dependent manner with 11 concentrations using 1:2 dilution starting from 10 µM.

The final assay concentrations were 0.05 mM Kynuramine and 0.1 U/well MAO A or 0.125 U/well MAO B, respectively in a potassium phosphate buffer (0.1 M) at pH 7.4 made isotonic with KCl.

In a pre-screen at 100 µM concentration for MAO A 145 compounds and for MAO B 108 compounds were identified that reduced enzyme activity to < 20 %. Subsequently, IC_{50} values were determined (see Supplementary Table 3).

All measurements were carried out at least in triplicate. All data were evaluated using Microsoft Excel and IDBS XLfit. From the kinetics data over 30 minutes, a linear fit over at least 10 minutes yielded the fluorescence increase proportional to the reaction velocity. These rates were then normalized to the positive (no inhibitor, 100% activity) and negative (no enzyme, 0% activity) controls. The reference compound served as additional quality control. For the concentration-dependent measurements, the corresponding rates were fitted together with the compound concentrations using a four-point logarithmic fit. The resulting IC_{50} values as well as their standard deviation are reported.

1.6.5 Sphingomyelinase assays

This assay was performed by Dipl.-Chem. A. Roth in the group of Prof. Dr. C. Arenz at the Humboldt University of Berlin. Parts of the procedures were developed and supplied by A. Yektaoglu (M.Sc.) from the group of Prof. Dr. A. Giannis at the University of Leipzig.

Isolation of Sphingomyelinase

The isolation and purification was performed as described by Wascholowski and Giannis.[119]

Isolation of neutral sphingomyelinase containing microsomes:

One rat brain (stripped rat brains, Pel-Freez Biologicals) was homogenized with a Teflon pestle in 10mL of 50mM Tris/HCl buffer (pH 7.4), containing 2 mM EDTA, 5 mM EGTA, 5 mM DTT, 5 mM ß-Mercaptoethanol and a protease inhibitor mix after washing with isotonic sodium chloride solution. The crude homogenate was centrifuged (1000 x g, 15 min) to remove debris. The supernatant was centrifuged at 100,000 x g, 90min, and the pellet was solubilized in homogenisation buffer containing 1.7% (v/v) Triton X-100. After rocking at 4°C for 2 hours, the microsome preparation was stored at -80°C or loaded onto a strong anion exchange column (Mono Q 5/50 GL, GE Healthcare).

Isolation of acid sphingomyelinase

One rat brain (stripped rat brains, Pel-Freez Biologicals) was homogenized in 10mL of 100 mM sodium acetate buffer (pH 5.0), containing 0.1% (v/v) Triton X-100. The brain tissues were homogenized for 3 rounds of 10 passes each. The crude homogenate was then centrifuged to remove debris (1000 x g, 15 min). The supernatant was centrifuged at 35,000 x g for 45 min and was filtered trough a sterile filter (ø 0.45 µm). The enzyme preparation was stored at -80°C or loaded onto a strong anion exchange column (Mono Q 5/50 GL, GE Healthcare).

Purification of Sphingomyelinase

A mono Q 5/50 GL column was flushed with 5 column volumes (CV) of equilibration buffer (nSMase: 20 mM Tris/HCl, 1 mM EDTA, 1 mM EGTA, protease inhibitor mix, pH 7.4; aSMase: 50mM sodium acetate, 1 mM EDTA, 1 mM EGTA, protease inhibitor mix, pH 5.0). After sample loading, the column was washed with 10 CV of equilibration buffer, followed by a linear gradient (3 CV) of 0 – 100% washing buffer (equilibration buffer, containing 1M sodium chloride) and maintained for 8 CV at 100% washing buffer. The final elution was carried out with a 10 CV - gradient from 0 to 1.1% (v/v) Triton X-100 in washing buffer.

Micellar nSMase Assay

The micellar nSMase-assay was performed by analogy to the method of M. Kölzer *et al*.[120]
The assays were performed in the presence of 200 mM Tris/HCl (pH 7.4) containing 5 mM magnesium chloride and 0.1% (v/v) Triton X-100. 40µL of the ^{14}C-labeled sphingomyelin (SM) (final SM concentration 100 µM), 7 µL of the enzyme preparation and 3 µL assay buffer (or inhibitor stock) were incubated for 30 min at 37°C. Inhibitors were added in the concentrations indicated in the graphs. The reaction was stopped by adding 800 µL chloroform/methanol 2:1 (v/v) and additional 250 µL water. The samples were extracted for 5 min and centrifuged at 13.400 rpm for 2 min. Radioactivity of released ^{14}C-phosphorylcholine in the aqueous phase was determined in a liquid scintillation counter (Packard TriCarb 2800Tr.).

Fluorescent aSMase Assay

The standard acid sphingomyelinase assay was performed in a 384-well-plate. Reaction mixtures consisted of 13.3 µL HMU-PC-substrate (HMU-PC 6-Hexadecanoylamino-4-methylumbelliferylphosphorylcholine), 13.3 µL reaction-buffer (100mM sodium acetate (pH 5.2), 0.2% (w/v) Na-TC, 0.02% (w/v) NaN_3, 0.2% (v/v) Triton X-100) and 13.3 µL enzyme preparation. Inhibitors were added in the concentrations indicated in the graphs. The reactions were incubated for up to 3 hours at 37 °C in a plate reader (FLUOstar OPTIMA, BMG labtech) and the fluorescence of HMU (6-Hexadecanoyl-4-methylumbelliferone) was measured (excitation 380 nm, emission 460 nm) in real time.

1.6.6 STAT protein assay

The STAT assay and the Western blot experiments were carried out by B. Sperl from the group of Dr. T. Berg at the Max Planck Institute of Biochemistry in Munich.

Fluorescence Polarization Assays for STAT Proteins

The final concentration of buffer components used for all fluorescence polarization assays was 10 mM Hepes (pH 7.5), 1 mM EDTA, 0.1% Nonidet P-40, 50 mM NaCl, 1 mM DTT, and 10% DMSO. The sequences of the peptides were: STAT1: 5-carboxyfluorescein-GY(PO_3H_2)DKPHVL;[143] STAT3: 5-carboxyfluorescein-GY(PO_3H_2)LPQTV-NH_2;[144] STAT5: 5-carboxyfluorescein-GY(PO_3H_2)LVLDKW;[145] Lck: 5-carboxyfluorescein-GY(PO_3H_2)EEIP.[144] The design of expression plasmids coding for STAT1, STAT3, STAT5b, and Lck as well as their expression and purification has been described.[143,144]

Proteins were used at the following final concentrations: STAT1: 60 nM; STAT3: 220 nM; STAT5b: 120 nM; Lck: 75 nM. Proteins were incubated with test compounds in Eppendorf tubes at 22 °C for 60 minutes prior to addition of the respective 5-carboxyfluorescein labeled peptides (final concentration: 10 nM). Fluorescence polarization was analyzed after an additional incubation at room temperature for 60 minutes. Binding curves and inhibition curves were fitted using SigmaPlot (SPSS Science Software). All competition curves were repeated three times in independent experiments.

The known STAT5 inhibitor **REF**[110] (see Figure 4a) was included to allow for direct comparison of compound activities, and an anti-screen for the binding to the SH2 domain of Lck to exclude that they are general SH2 binders. Both hit compounds did not bind to the SH2 domain of Lck.

Western Blot Analysis with pSTAT Antibody

The Western blot analysis was performed as described by Berg *et al*.[110] Daudi cells (ATCC number CCL-213) were grown in RPMI 1640 media according to ATCC recommendations, incubated for 1 hr with the indicated concentration of compounds, and stimulated for 5 min with 5,000 U of interferon-α_{2A}. The final DMSO concentration for the test compounds and the controls was 0.4%. Cells were washed twice with ice-cold PBS and lysed with 80 µL of buffer (50 mM Hepes, pH 7.5, 150 mM NaCl, 1 mM EDTA, 10% glycerine, 1% Triton X-100, 10 mM $Na_4P_2O_7$, 10 mM NaF, 20 µM Na_3VO_4, 10 µM PMSF and 100 ng/ml aprotinin). Proteins were separated by SDS-PAGE, transferred to nitrocellulose, and probed with the relevant antibodies (anti-STAT5 pTyr694, anti-STAT3 pTyr705, anti-STAT1 pTyr701, anti-STAT5, anti-STAT1 (all rabbit) from Cell Signaling; rabbit anti-STAT3 from Santa Cruz; rabbit anti-actin from Sigma). Secondary horseradish peroxidase (HRP)-conjugated antibodies were from DakoCytomation.

2 Protein Structure Similarity

2.1 Introduction

2.1.1 Protein structure and its application in small molecule ligand design

Proteins are modular entities that consist of several different modules, so called domains. Although there is no clear-cut definition for the term “domain” it may be described as an autonomous folding unit consisting of a single peptide chain.[146] The term “fold” describes the spatial arrangement of secondary structure elements like α-helices and β-sheets, relative to one another.[147] Modern structural biology combined with bioinformatics analyses of genome-based data have revealed that protein folds are well conserved in Nature and during evolution.[148-152] The SCOP (Structural Classification of Proteins) database[132,153,154] predicts about 1,200 folds corresponding to 38,000 entries in the protein data bank (PDB).[155] Different fold comparison methods disagree on the total number of folds, and depending on the algorithms used, the estimates of the overall number of folds range from 1,000 to 10,000 which is very low considering the thinkable possibilities.[152] Another feature of the conservatism in protein domain folding is that the distribution of folds is highly non-homogeneous with some folds occurring abundantly and some rarely.[156-158] It has been proposed that a majority of protein domains can be attributed to ca. 1,000 most commonly observed folds. Based on this structural conservatism in protein architecture Protein Structure Similarity Clustering (PSSC) was developed as a guiding principle for the selection of biologically validated starting points for compound library development.[159-163]

In the protein world, structural conservatism and diversity combine on two different levels: conservatism in the more macroscopic, i.e., the structural level, and diversity on the microscopic level, i. e. the individual amino acid sequence. The fold defines the scaffold of the protein, i.e., the 3D structure of the amino acid backbone, as well as the shape and size of the active site and the spatial orientation of the catalytic residues. The individual amino acid side chains forming the active site and its catalytic residues determine the molecular interactions between the protein and the ligand. The same fold can be assembled by amino acid sequences with as little as a few percent sequence similarity. Thus both, fold and sequence, together determine the binding properties of any protein and enable the vast number of specific functions to be carried out by a limited number of fold types.[164-167]

Protein structure is applied in the discovery and development of small molecule modulators of protein function in many ways and at different levels of resolution. Atomic resolution, that is the individual positions of atoms and functional groups, is used among others in molecular docking[168-172] where an ensemble of conformations of a ligand structure is docked into a protein structure in a computer. All solutions are then ranked according to scoring functions thereby

generating hypotheses on the potential binding mode of the molecule and relative potencies.[173-178] One of the most extensive assessments of docking algorithms and scoring functions was published by Warren *et al*.[179] Other approaches using high quality protein structures are programs for *de novo* ligand design where in many cases ligand structures are assembled in a given binding site. The methods involve 'molecular growth' by addition of atoms and functional groups, sometimes also according to reaction databases and virtual building block collections, or fragment linking where known inhibitors and drugs are fragmented first and these fragments recombined in a second step.[180-184] Another approach to exploit the complementarity between ligand- and protein-based interaction points are the fingerprints for ligands and proteins (FLAP).[185] This method is based on a pharmacophore-like representation of the molecules where hydrophilic and hydrophobic patches or hydrogen bond donor/acceptor regions are mapped and combined into a 3D fingerprint. Such fingerprints can be generated for ligand binding sites as well as for small molecules. A relatively simple similarity calculation between the fingerprint of a given protein binding site and the fingerprints of a set of small molecules can identify the best matches that can be expected to bind the protein better than the rest of the data set.

Cavbase is an extension to the Relibase database of protein structures and groups proteins according to the similarity of the shape of their binding site. The resulting shape similarity clusters can also be explored for ligand design since cavities with similar shape and surface potential are prone to bind similar ligands.[186-189]

2.1.2 Protein structure similarity clustering (PSSC)

The idea of protein structure similarity clustering (PSSC)[43] is based on the structural complementarity beween proteins and their small molecule ligands that forms the basis for binding interactions, e.g., hydrogen bonds or hydrophobic interaction, and, ultimately, for modulation of protein function. This structural complementarity can be addressed at different levels of detail, for instance, the atomic level, the fold level etc.

As explained above, many approaches, e.g. docking, aim at predicting binding based on atomic resolution, for example, the interactions between individual atoms and functional groups. Although such a detailed analysis can be expected to yield results referring to individual molecules, the precision of these results largely depends on the underlying algorithms and approximations as well as on the data like protein crystal structures whose resolutions can greatly vary. The use of homology models, that is, structural models of proteins that are built on the known structures of proteins sharing a high sequence similarity with the protein of interest, in docking often leads to less precise predictions.

PSSC does not address complementarity of small molecules and proteins structures at the atomic level but rather at the more abstract level of scaffold structures.[190] In chemical structures, a scaffold is defined as the (often cyclic and rigid) core structure to which the

substituents are attached. The scaffold provides 3-dimensional pre-organization of the side chains that form the majority of the molecular interactions. By analogy to this chemistry-centred view, proteins can be divided into a scaffold and its substituents as well. In a protein, the backbone comprising of the amide bonds and the amino acid carbon between them, also called the c_α trace, forms the tertiary structure; the fold of the protein. Although these atoms can not fold into the protein tertiary structure without the side chains, they are nonetheless found to provide a spatially defined scaffold to which the side chains forming most of the molecular interactions are attached.

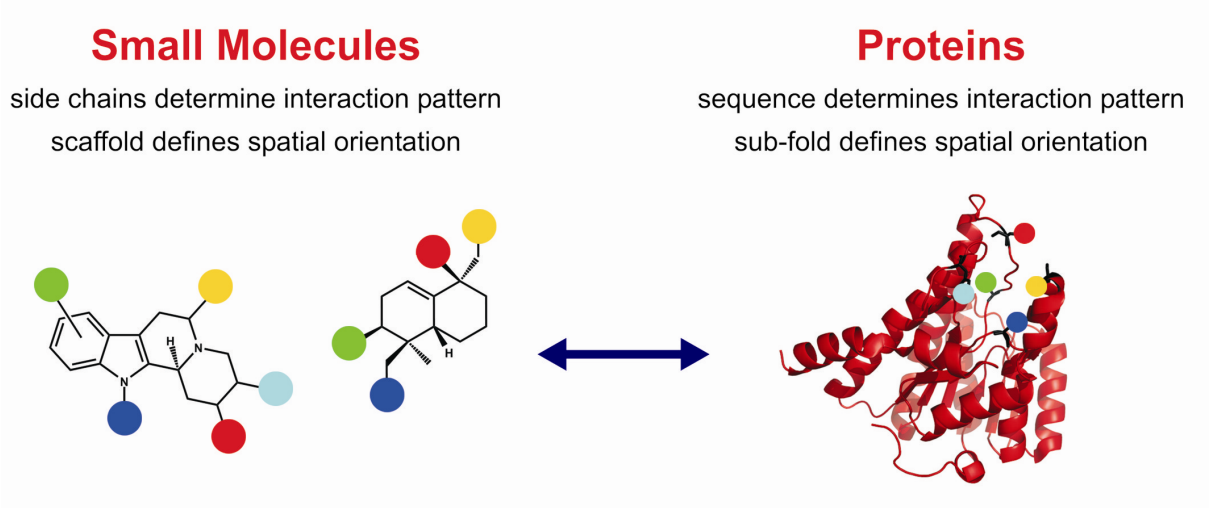


Figure 29: Scaffold-substiuent analogy between small molecules and proteins. The scaffolds in small molecules orient the substituents marked as coloured circles forming the interactions. In proteins, the sub-fold defines the spatial arrangement and orientation of the side chains. Binding occurs when compatible substituents (circles with the same colour) match in their spatial position so that they can interact.

The basic hypothesis of PSSC is that for successful binding and, hence, modulation of protein function, complementarity between the chemical scaffold and the protein scaffold is required (see Figure 29). If this hypothesis is correct, then the converse argument should also hold and proteins that exhibit similar 3-dimensional structures of the binding site should bind similar molecules. Or, in other words, if proteins are grouped for similar sub-folds (= scaffolds) around the binding site, then these clusters should bind molecules embodying similar or even the same scaffolds. It is important to note however, that PSSC does not predict binding of individual molecules; it exclusively matches protein and small molecule scaffold families.

As laid out in sub-section 1.1.1 the number of scaffolds found in synthetic molecules[7] and natural products[191] is fairly limited. However, any given scaffold type can be decorated with a large number of diverse substituents at different positions theoretically enumerating a large number of possible molecules. Analogous to the protein fold, the scaffolds define the

frameworks of the protein ligands while the individual substituents decorating the scaffolds define the molecular interactions between the individual ligand and its target.

Even with similar backbone structures, proteins can display very diverse interaction patterns due to different amino acid sequences. Therefore, it is necessary to develop compound collections because only compound collections can provide sufficient chemical diversity in the small molecules to match these different interaction patterns in the protein. Most likely, a particular natural product will not bind to all structurally similar binding sites found in different proteins. One can imagine an extreme example of two nearly identical binding sites, one of which carries a positively charged residue and the other one a negatively charged amino acid side chain at a similar position in their binding pocket. A negatively charged ligand would in the first case be bound tightly through a salt bridge and in the other case be repelled from the pocket. Consequently, it is necessary to generate sufficient chemical diversity to match biological diversity in the quest for biologically active molecules/ligands for proteins.

In practice, the PSSC procedure (see Figure 30) started with a search of the full structure of a protein of interest to find structurally similar proteins using the Dali database on protein structure families (FSSP) and Combinatorial Extension (CE) algorithm.[192-194] The searches were performed across the entire Protein Data Bank (PDB) and yielded lists of structurally similar proteins ordered by decreasing similarity. The entries that were considered as interesting according to different criteria, e.g., pharmaceutical relevance or low sequence similarity, were then inspected manually. For this step, ligand-sensing cores, i.e., spherical cut-outs of the proteins structure centred on the binding site of interest were manually isolated and aligned. The cores that showed sufficient similarity in their 3D structures were assigned to a protein structure similarity cluster. Known ligand types for one individual cluster member are regarded as complementary to this sub-fold and thus used as a template for the development of compound collections targeting all cluster members. The manual procedure described was successfully applied in the discovery of novel 11β-hydroxysteroid dehydrogenase 1 inhibitors.[43]

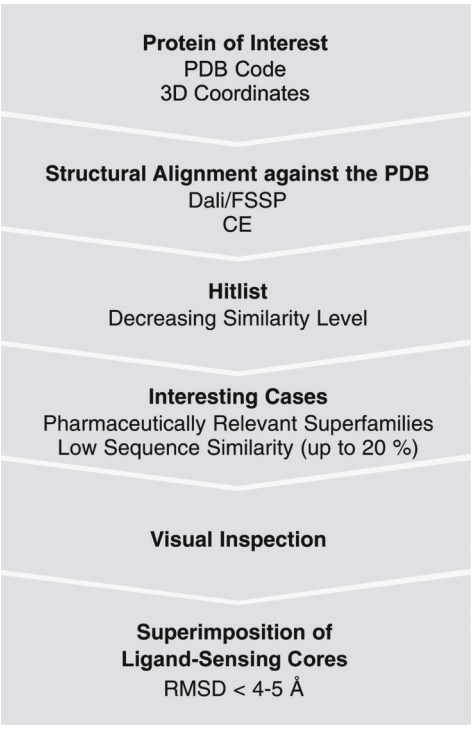


Figure 30: Flow-chart of the manual PSSC procedure. The structure comparisons yield hitlists which are manually analyzed for interesting cases. Subsequent visual inspection and alignment of ligand-sensing cores verifies the structural similarity.

2.2 Aims

As shown in the introduction, a manual workflow for PSSC had been developed by Koch *et al*. and successfully applied in a first application example.[43] To obtain a broader proof-of-concept the analysis of a larger number of protein structures from different classes of proteins would be required. Since the initial procedure involves considerable manual work on ligand sensing core extraction, structural alignments and reviewing results only limited datasets can be processed. The parameters underlying the analysis, for example the definition of similarity, have to be chosen by the examiner, which may lead to non-standardized results. Therefore, the aim was to automate and standardize the entire PSSC analysis process, as well as to validate the similarity parameters used for the clustering. In detail, the aims of this project were as follows:

- Design and implementation of an automated or semi-automated PSSC process for the analysis of large protein structures sets. This includes the modules for ligand-sensing core extraction, structure comparison, clustering and final evaluation of the clusters.

- Evaluation of a medium sized set of structures to validate the automated process. This would include cross-checking with established databases for structural family relationships such as the structural classification of proteins (SCOP) or the CATH database.

- Experimental validation of a PSSC cluster by screening of a compound library embodying the scaffold of a known inhibitor of one cluster member against one or more of the other proteins in the cluster.

2.3 Results

2.3.1 Design and implementation of a fully automatic PSSC process

The amount of protein structure data available from modern protein crystallography technology has been fast growing over the last two decades (see Figure 31). Any analysis capturing a

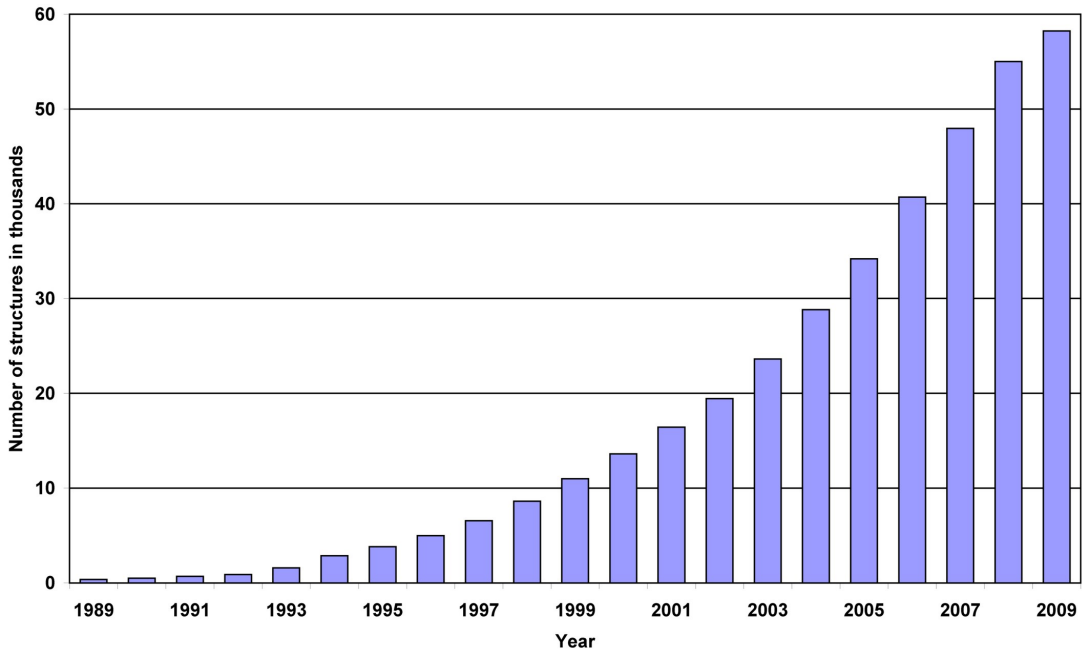


Figure 31: Total number of protein structures stored in the PDB over the past two decades illustrating the exponential growth.

representative number of structures to obtain a proof-of-concept demands the development of a semi- or fully automated process for structural similarity clustering. Such a process can be broken down into several functional modules dealing with the ligand-sensing core extraction, the structural alignments, scoring of the alignments themselves and subsequent clustering and manual analysis (see Figure 32). There is a number of issues to be considered in the design and implementation of the overall process as well as of each of these modules.

Figure 32: Functional modules of an envisioned automated PSSC process.

The design of the overall process was based on the lessons learnt from the manual PSSC analyses. One of the most time-consuming steps in the manual approach is the analysis of the ranked hit list resulting from the FSSP database query since it contains many false-positive alignments. This is due to the alignments of entire protein structures where similarity may be

detected in domains or regions remote from the binding site of interest. Only in the last step are the decisive parts of the structure, the ligand-sensing cores centred on the binding site, manually extracted from the protein structures and their structural alignments are visually inspected. To avoid too many false positives in the analyses, the first step of an automated PSSC procedure would be the generation of the ligand-sensing cores removing all parts of the protein that are not of interest for the analysis. The ligand-sensing cores will then be structurally aligned with one another in the next step. In the manual procedure, structural similarity online databases were queried such as the FSSP[192,195] and the Combinatorial Extension (CE)[194] databases, comprising of structural alignments for the protein structures contained in the PDB that were pre-computed using the Dali[196] and CE[193] method, respectively. Since these databases only contain alignments computed on at least one chain or domain of each protein, they could not be used to determine the structural similarity between the ligand-sensing cores. In the FSSP database, “close homologs” of proteins are clustered together if they exhibit a sequence similarity of more than 70% and structural alignments are only carried out between cluster centres.[197] This restriction is a relict from the early nineties when computational power was very limited and 11 MByte of storage space occupied by the first FSSP version was quite enormous at the time. Because of the substantial increase in the number of protein structures (see Figure 31) and the high redundancy found in the PDB[198,199], the clusters grew very large and the representative of each cluster is subject to constant change, which reduces the reproducibility of results over time. For these reasons, structural alignments of all ligand-sensing cores against all have to be calculated locally. The choice of the program/algorithm as well as the implementation is discussed later. The assessment of the structural similarity between two structures, i.e., the score, is an important aspect in the whole analysis. Several programs use different measures of similarity and for each measure, the minimal similarity for two proteins to be clustered together needs to be determined. In the manual PSSC procedure, clusters were selected according to their similarity to the cluster seed, as well as to their pharmaceutical relevance. Manual clustering of analyses with hundreds or even thousands of ligand-sensing cores is clearly impossible. Moreover, the selection according to ‘soft’ criteria such as pharmaceutical relevance may introduce a bias into the whole analysis. Therefore, the resulting lists of structural similarity will be transferred into a database and subsequently clustered by an automated clustering algorithm. At the last step of each analysis, a visual inspection and in-depth analysis of the clusters considered interesting for further research is conducted.

Extraction of ligand-sensing cores

Before the design and implementation of the processes for their extraction, ligand-sensing cores need to be defined in an objective way. Such a definition must include the size of the cores as well as their centres and most importantly, also the spatial location of the binding site that forms

the centre of the ligand-sensing core. To obtain such a definition proved to be difficult, however, a database of catalytic residues of enzymes was eventually obtained. The so-called 'Catalytic Site Atlas' (CSA) contained more than 15,000 entries of catalytic sites extracted from the literature and sequence alignments in its first version.[200] The latest version 2.2.10 from October 2008 contains annotation to 23,265 proteins based on almost 1,000 literature references.[201] To determine a sensible value for the radius of ligand-sensing cores, first the size distribution of 8444 catalytic sites stored in the Catalytic Site Atlas was determined. To this end, the centre of mass was calculated based on the alpha carbon atoms of the catalytic residues and maximal radius between the centre of mass and the residues comprising the catalytic site determined. It was found that more than 50% sites have a radius between 4 and 7 Å whereas 97.4% of the sites are smaller than 15 Å (see Figure 33). From this result, a minimal radius of the ligand-sensing cores of 10-15 Å was deduced, depending on the remaining procedure.

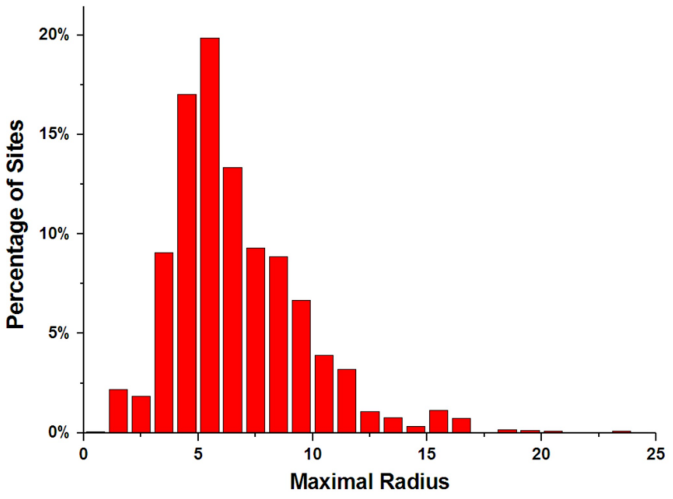


Figure 33: Size distribution of the catalytic sites in the Catalytic Site Atlas.

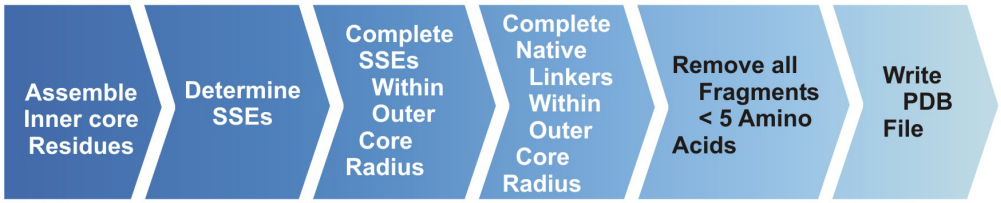


Figure 34: Functional modules of the final ligand-sensing core extraction process.

With the information about the catalytic residues and the radius, in principle ligand-sensing cores can be constructed. However, initial tests resulted in cores containing many structural fragments, often even fragmented secondary structure elements (SSEs). Since PSSC builds on comparisons of the sub-fold around the binding site, that is the spatial arrangement of secondary structure elements, structural similarity could be blurred by incomplete SSEs. Therefore, a multi-step process was designed during which more complete, less fragmented cores are built (see Figure 34). To achieve this, an additional 'growth zone' between the inner core radius and a larger outer core radius was defined. SSEs and linkers extending into this zone will be completed, thereby generating more intact and less fragmented sub-folds. For an

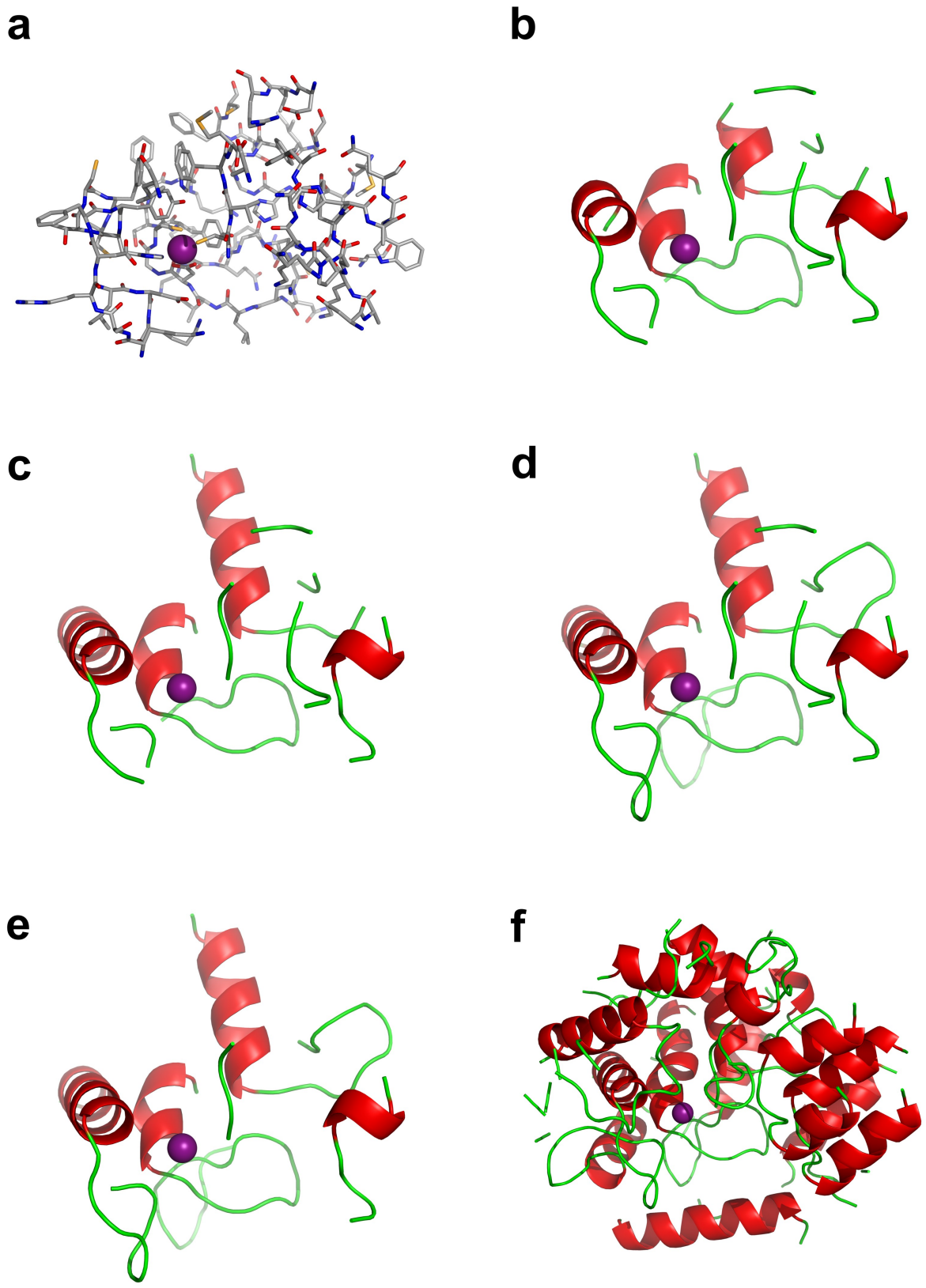


Figure 35: Ligand-sensing core of rat farnesyltransferase (1JCR) at each stage of the ligand-sensing core extraction process with an inner core radius of 15 Å and an outer core radius of 25 Å, resprectively. The zinc atom in the catalytic site is shown as purple sphere. a) Residues in the inner core radius. b) SSEs formed by the residues within the inner core radius. c) Core with completed SSEs reaching into the growth zone. d) Core with completed linkers. e) Final ligand-sensing core where small fragments forming no SSE were removed. f) Secondary structure of all residues within the outer core radius. The difference in size and number of secondary structure elements to the 'grown' ligand-sensing core in e) is obvious.

inner core radius of 15 Å, the optimal outer core radius was determined to be 25 Å (for details, see Methods section 2.6.1). The core extraction process starts with the extraction of all residues within the inner core radius of 15 Å (see Figure 35a). The centre of the core is the calculated centre of mass of the alpha carbons of the catalytic residues defined for each site in the CSA. In the second step, the SSEs are determined for the full structure using the Dictionary of Protein Secondary Structure (DSSP)[202], for a long time the gold standard in secondary structure recognition (see Figure 35b). Since the initial ligand-sensing cores can contain fragmented SSEs that would not be recognized by the DSSP program, the full protein structure is used in this step. Thus, pre-computing a set of DSSP files for the PDB that is then used for ligand-sensing core extraction saves computational resources and time. The program for ligand-sensing core extraction checks for a pre-computed DSSP file in a specified directory and generates only the missing files. The SSEs whose origin lies within the inner core radius and that stretch into the outer radius are completed. This means that the missing residues belonging to the SSE are added from the PDB file as long as the SSE ends within the outer radius (see Figure 35c). All elements (including linkers as described in the next step) that are added during the process are genuinely taken from the protein structure; no artificial elements are created and introduced. The resulting core exhibits a more intact sub-fold than the initially created core. In the next step, residues are added that link two chains in the ligand-sensing core and that are located within the growth zone (see Figure 35d). This step reduces the number of fragments in the core and is important because more fragments or individual chains lead to more complicated alignments and can, eventually, induce complete failure of some alignment algorithms. This is also the reason why all fragments comprising of less than 5 amino acids and not forming a secondary structure element, that is, a helix or a β-sheet, were removed. With this final step, the ligand-sensing core is completed (see Figure 35e). It exhibits a well-defined sub-fold comprising of a few SSEs in a particular spatial arrangement. In comparison to a core incorporating all residues within the outer core radius (see Figure 35f) the ligand-sensing core generated by the described procedure is much more focused on the binding site, which is one key requirement for successful application of PSSC. The core extraction procedure and the need for its complexity are discussed in sub-section 2.4.1. Its function described in pseudo-code can be found in the Experimental section 2.6.1. Application of this core extraction process to a large data set is described in sub-section 2.3.4.

Structural alignments

Once a set of ligand-sensing cores is generated, the next step is the computation of the structural similarity of these cores. In PSSC, structural similarity is defined on the tertiary structure or sub-fold level meaning the spatial arrangement of secondary structure elements in relation to one another. The field of structural comparison of proteins has been extensively

reviewed[203-205]. Over the past years, a plethora of algorithms for protein structure comparison have been developed[193,196,206-218]. Hence, usage of an established algorithm seems much more advantageous compared to the development of a new algorithm. For the choice of a suitable algorithm several criteria were developed, which include the availability of a useable offline computer program, good performance (fast as well as accurate), measures for structural similarity and significance of the alignment and, if possible, generation of rotation-translation matrices[5] for superimposition of the two structures which is helpful for visual inspection. Many of the tools cited are not available as stand-alone programs or have not been applied extensively enough to judge their quality. After extensive searches, three programs employing different alignment algorithms were investigated more closely because they were accepted as standard methods in the files: Dali[196,219,220], CE[193,194] and the vector alignment search tool (VAST).[206,215,216] Whereas Dali and CE were available as offline programs, VAST only offers an online search through a webpage.[221] Submitting structure files manually to a website for computation of the structural comparison may be feasible for smaller analyses. However, for the comparison of large scale data sets, an offline program is required to be used in an automated alignment process. Therefore, VAST was excluded from the list. The performance of several protein structure alignment methods including Dali and CE were compared by Sierk and Pearson in 2003 based on extensive testing.[222] The authors observe that Dali produces the best quality alignments. Dali is also well-known and widely accepted, and is available free as a stand-alone computer program.[219,220,223] Moreover, it generates a root mean square deviation (RMSD) value as a quality criterion for the alignment as well as a Z-Score indicating the significance. The importance of both values will be described below. Rotation-translation matrices are routinely generated and can easily be used to create superimpositions for visual inspection. Moreover, since the previous manual PSSC analyses were performed using the FSSP database[192,195] built with the Dali algorithm[196], the use of Dali allows the direct comparison of results, e.g., for validation purposes. Following the argumentation laid out above, it was decided to use DaliLite for the structural alignments.

Dali returns two values for each structural alignment: the root mean square deviation (RMSD) and the Z-Score. The RMSD value is a measure for the similarity of two structures: the lower the RMSD, the higher the similarity of the structures. Usually, RMSD values of 1.0 or less are considered to be of identical structures. The RMSD

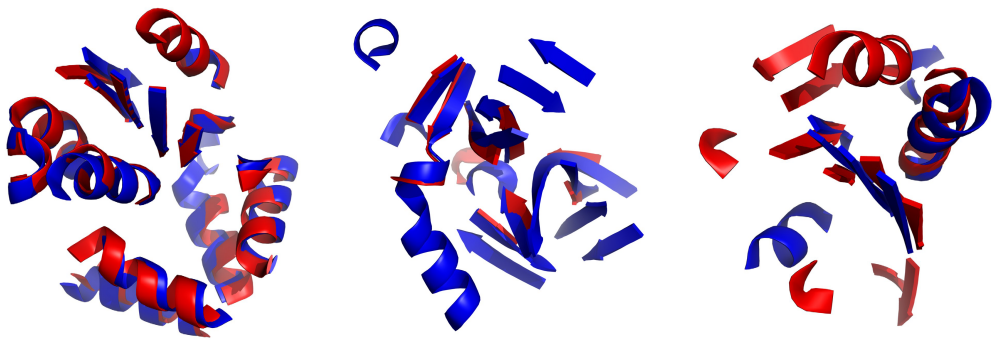
$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} d_n^2} \text{ [Å]}$$

Figure 36: Formula for calculation of the RMSD value.

[5] Rotation-translation matrices describe the mathematical transformation of the coordinates of one structure to superimpose it with a second structure. It usually consists of a 3x3 matrix **M** representing the rotation of the protein in space as well as a 3-dimensional vector **v** describing the translation of the structure. Calculation of the superimposed 3-d coordinate vector **c'** for each atom is possible by a simple mathematical transformation: **c' = cM + v**.

is calculated by averaging over all distances between the structurally equivalent atoms in the two structures (see Figure 36). It is an overall measure that averages similarity over the whole aligned sequences. If the reasonably similar atoms for RMSD calculation are not determined, it is possible to obtain a higher RMSD value despite some very similar substructures – if the remaining structure parts that are used in the RMSD calculation are highly dissimilar. Therefore, Dali determines the structurally aligned atoms first and uses only these in the RMSD calculation. In small structures like ligand-sensing cores, almost all residues should be structurally aligned.

The Z-Score indicates the statistical significance of a structural alignment and was found to be the most reliable measure for this purpose.[222] Such a measure is needed because there is a high probability that during any random superimposition of two protein structures, several atoms or even residues of both structures will end up in the same space by chance. Such random superimpositions will give lower RMSD values but do not indicate structural similarity. For



Z-Score	27.1	13.1	2.3
RMSD [Å]	1.6	1.6	2.9
Alignment Length	182	113	58
Sequence identity [%]	40	38	9

Figure 37: Three examples of structural alignments illustrating the importance of the Z-Factor. In all alignments, one structure is coloured in red and the other on in blue. The structural similarity visibly decreases from left to right; the RMSD does not change much whereas the Z-Score does.

illustration, three cases with different Z-Score values are shown in Figure 37 illustrating the decreasing structural similarity indicated by the declining Z-Score. The example on the left hand side shows two perfectly aligned structures as indicated by a very high Z-Score of 27.1 and an RMSD of 1.6 Å. The two structures shown in the middle do align reasonably well but less so than the first example. The β-sheets do not overlay exactly and some of them as well as the α-helix in the blue structure do not have counterparts in the red structure. Notably, the RMSD is the same as in the first example, 1.6 Å. Only the decrease in Z-Score by 50% indicates that although some structural similarity is present, the two structures are less similar than in the first

example. The alignment on the right hand side looks even less similar: many SSEs do not have any counterpart and those that do (one α-helix and the central β-sheets) do not overlay well. Although the RMSD rises moderately to 2.9 Å the Z-Score drops by a factor of six from the example in the middle. In general, a Z-Score smaller than two indicates that the structures only randomly overlay but that there is no significant structural similarity. The examples given in Figure 37 illustrate the need for a measure reflecting the statistical significance of the structural alignments. Without the Z-Score, a low RMSD value can indicate good alignments and structural similarity even though it may be just a random superimposition. The theory behind it is, non-mathematically speaking, as follows: if two protein structures are similar to each other, an optimal alignment can be found where the RMSD value is minimal, i.e., both structure deviate as little as possible from each other. Since both cores are structurally similar and ideally aligned, any small movement of one core, either rotation or translation, will impar the alignment of many residues at once and thereby significantly increase the RMSD value. The larger the increase of the RMSD value, the better and more significant the structural alignment. If, by contrast, only some atoms randomly superimpose, any movement of one protein will not change the RMSD much because for the atoms that no longer superimpose, other atoms randomly will. To determine the Z-Score, a number of small movements of one structure are carried out and the resulting change of the RMSD value is stored. The Z-Score is then calculated from the average RMSD change over all these movements.

The process for computation of the Dali structural alignments is illustrated in Figure 38. The various steps shown in this flow chart as well as the reasoning for the process design shown are explained in the following paragraphs.

Since calculation of Dali alignments takes several seconds per pairwise alignment, the computation of all-against-all alignments of large data sets was distributed over multiple processors of an in-house Linux cluster. This was achieved using a self-written perl script that distributes the Dali program as well as the ligand-sensing core set over the individual nodes together with a pre-computed DSSP dataset for all ligand-sensing cores. Dali uses DSSP to identify the secondary structure elements that are then aligned structurally. The pairwise alignments are then calculated, subsequently generating four individual files that contain the result of the alignment. To avoid problems due to the automatic naming of the files (the result files would be overwritten with each alignment!), a folder hierarchy was created that consists of one folder for the template core that contained individual folders for all cores against which the template was aligned. This structure ensures that the data for each alignment reside in their own folder identified by the folder structure.

For the subsequent clustering as well as for evaluation it is beneficial to store all the data in a relational database system like MySQL. Therefore, a program was compiled in Java that processes all the files for each alignment, extracts the important data including the RMSD and

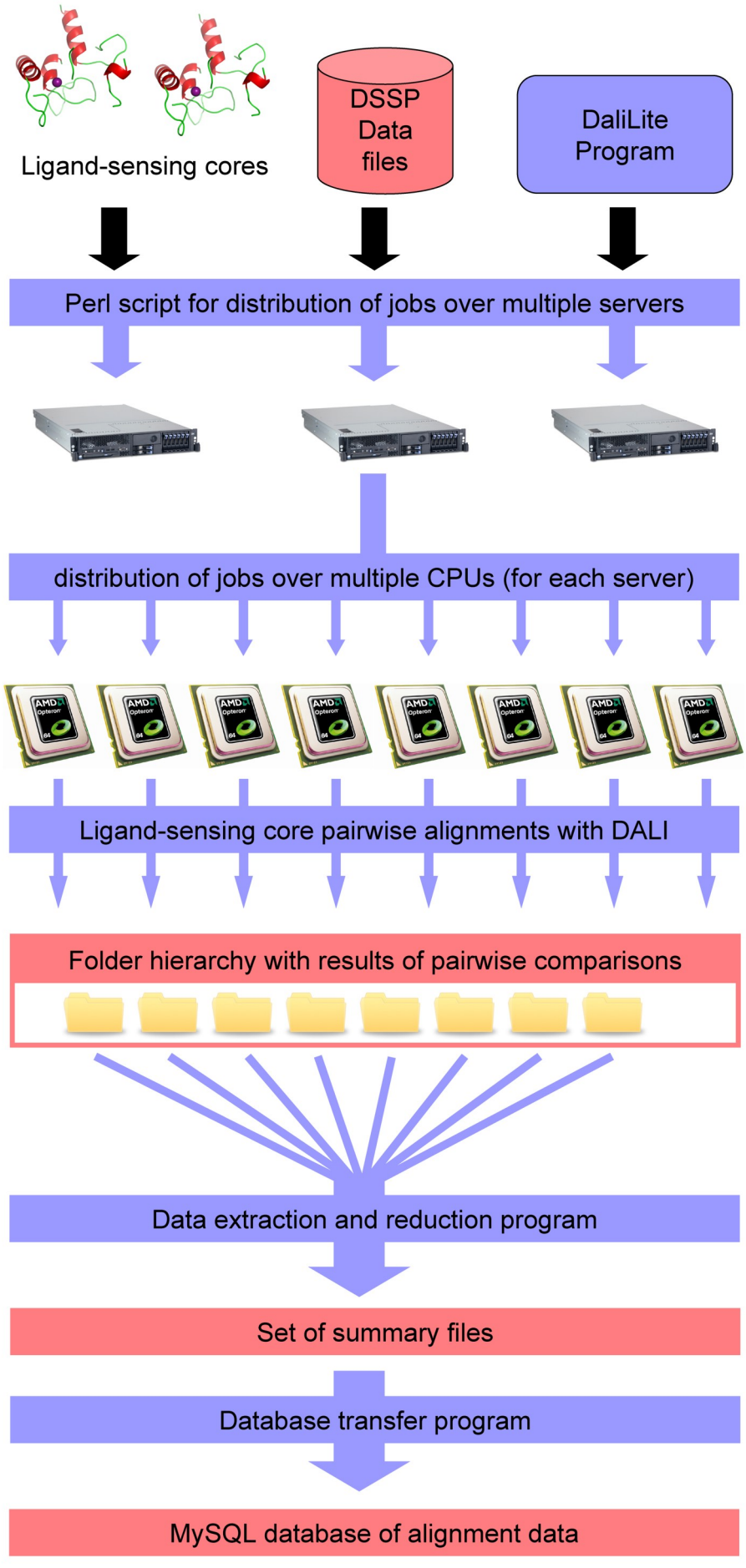


Figure 38: Flowchart of the distributed structural alignment process using Dali. The modules coloured in blue denote programs, those in red data.

the Z-Score, sequence identity, rotation-translation matrix and writes them into compressed data files. This intermediate step was introduced because the data processing was also distributed into a number of parallel jobs spread over multiple processors. Each job produces its own result summary files because in case of errors only the job where the error occurred needs to be re-run. Were all results directly to be written into a database, one would then need to verify that this particular result has not been stored prior to writing each record. Such a procedure would produce considerable network traffic and many database queries that would significantly slow down the data extraction process. Instead, another Java program was coded to extract the data from the result summary files and transfer them into the database. For technical details of the structural alignment process and the data processing, please see the Experimental sub-section 2.6.2. The structural similarity database is clustered in the next step by a custom programmed clustering algorithm.

Clustering of the results

"Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters)".[224] Naturally, each clustering is a compromise between two contrary trends: the homogeneity of clusters (the more clusters, the more homogeneous the individual clusters will be) and the classification aspect (the lower the number of cluster, the more effective the classification). From each cluster, a member is chosen as the representative, the so called 'cluster centre', that is the instance most similar to all other members of the cluster. Clustering is usually guided by a 'distance function' describing the dissimilarity between two entities. There are two criteria that need to be met by the distance function to be applied to clustering: first, the function needs to be continuous and to be able to describe the full range of occurring similarity. Second, the function needs to produce one value. Almost all clustering algorithms work with exactly one distance value. If there is no such value for some entities or if there are more than one, the clustering does not work.

It is these two criteria that prohibit statistical clustering based on the similarity output of the Dali algorithm. Whereas the RMSD value can, in theory, describe any degree of dissimilarity that is possible (up to infinity), Dali only calculates an RMSD value in cases where at least some atoms or residues superimpose. As described above, this protects against too many non-aligned atom or residue pairs distorting the RMSD value. The second criterion is not met as well since any clustering would need to take into account both values describing similarity, the RMSD and the Z-Score. Relying only on the RMSD value could lead to false results due to non significant, random alignments.

Therefore, another clustering method had to be developed taking into account both cases, the missing alignment (no RMSD value) and the Z-Score. Its design was based on the OptiSim algorithm developd by Clark.[225] This is a clustering algorithm that generates diverse

representative subsets from the given set of entities. It is, however, linear in nature and, thus, not reproducible. The clustering result can differ every time the algorithm is run if the number or sequence of the entities changes. This will become clearer during the explanation of the process.

The developed clustering process is shown in Figure 39. The list of ligand-sensing cores is processed sequentially, and for each core the algorithm verifies whether one or more cluster centres already exist that are closely resembling it. Although Holm and Sander gave a Z-Score of two as threshold for non-significant alignments, in the clustering decribed here, a value of four was implemented as the minimum Z-Score based on experience. For a comparison of the results borne from the clustering of the Catalytic Site Atlas data set, refer to sub-section 2.3.4. If the core is found to be similar to one or more cluster centres, it is assigned to the cluster with the most similar centre (see red path in Figure 39a). If no cluster centre is found that is sufficiently similar to the core, then this core forms a new cluster and becomes the centre of this cluster. Through this path, the first ligand-sensing core clustered automatically becomes the first cluster centre since no other cluster centres exist that could be similar to it. The devised clustering process was fully implemented in Java and used to cluster the Dali results from the MySQL database; for technical details, see Experimental sub-section 2.6.3.

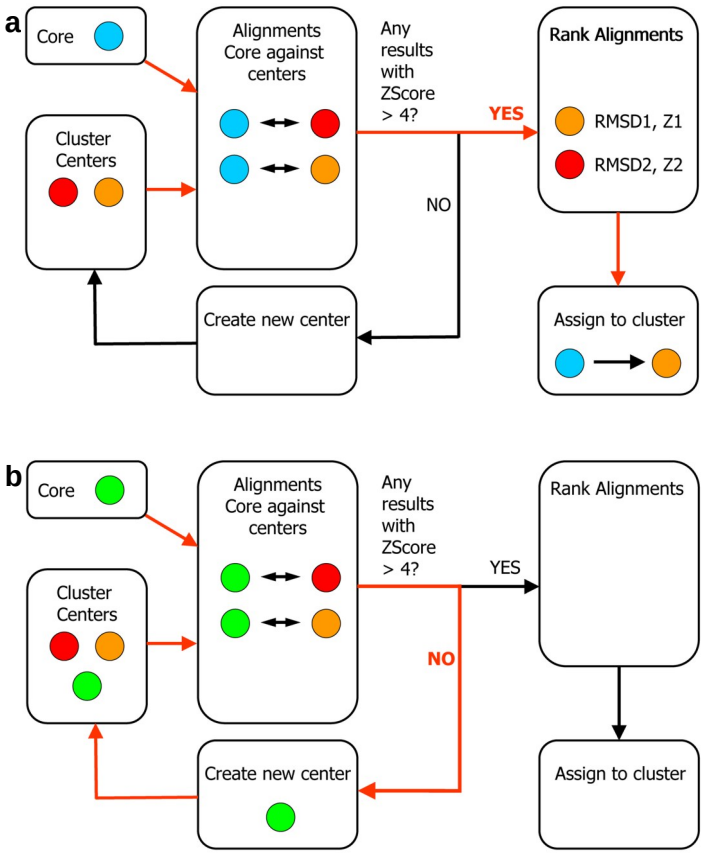


Figure 39: Diagrams describing the OptiSim-based clustering algorithm developed for PSSC. The red arrows mark the path of the algorithm for the given case. All the ligand-sensing cores in the set are processed sequentially. For each core, the program checks the similarity to all cluster centres already determined. Two cases can be differentiated: a) If one or more similar cluster centres are found, the core is assigned to the cluster whose centre it most closely resembles. b) If no similarity to any cluster centre is found, the ligand-sensing core will become the centre of a new cluster.

One drawback of this method is its large dependence on the order in which the cores are processed meaning that each entity gets assigned to the most similar cluster centre picked so far. This does not have to be the most similar centre at the end of the procedure, however.

This case is illustrated for three cores, A, B and C in Figure 40. A is not similar to C whereas B is similar to both but more to C than to A. If the cores are processed in the order A, B, C then A will become a cluster centre, B will be assigned to this cluster and C will form a cluster of its own. In this case, B was assigned to the cluster around A although it would be more similar to cluster centre B. Were the order of processing reversed, i.e., C, B, A then C would form the first cluster and B would be added correctly to the C cluster since it is more similar to C than to A, which will become the centre of the second cluster.

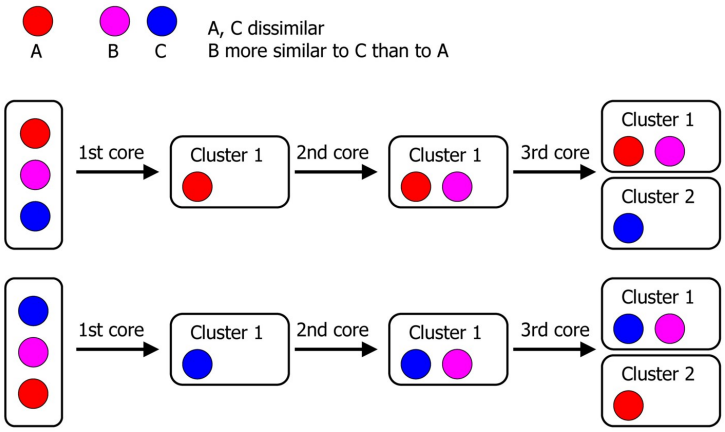


Figure 40: Illustration of the dependence of the clustering algorithm on the processing sequence of the ligand-sensing cores. Cores A and C are dissimilar whereas B is similar to both, A and C, but more similar to C than to A. If the cores are clustered in the sequence A, B, C then A forms the centre of a new cluster and B is added to this cluster. C forms the centre of a second cluster. However, if the order were to be reversed, C forms the centre of the first cluster to which B is then assigned and A starts a second cluster. In the first case, B ends up in a cluster in which it is less similar to the cluster centre than to the centre of cluster 2. In the second case, B ends up in the cluster where it is most similar to the centre.

In the many implementations of the OptiSim method, e.g., in Pipeline Pilot[226], this problem is circumvented by user input of an estimate of the number of clusters that should be generated. The algorithm then first picks as many cluster centres as described above and assigns the remaining entities to the clusters in a second step. However, such a procedure would be problematic in the PSSC case since there is no estimate on the number of clusters to be expected. One possible way to determine the optimal number of clusters could be to generate a number of N_1 clusters and assign each ligand-sensing core to one cluster. The average RMSD and Z-Score of each cluster are calculated and the distribution is saved. Then one starts over and generates N_2 clusters where $N_2 > N_1$. The larger the number of clusters, the more homogeneous the clusters become, that is, the more similar their members are to one another. This would be reflected in the average cluster RMSD and Z-Score distribution where the RMSD would be shifted to lower values as the Z-Score is shifted to higher ones. The optimal number of clusters would be reached as soon as the average cluster RMSD and Z-Score reach pre-defined values. However, especially for small numbers of clusters, the missing similarity values (RMSD and Z-Score) for non similar alignments would yield many singletons, i.e., entities that cannot be assigned to a cluster.

To overcome this drawback and also to speed up the similarity calculation process, a different solution was pursued, namely the use of protein fold fingerprints as described in the following

sub-section. A plethora of distance measures[227,228] can be applied with the fingerprints and many of them are amenable to statistical clustering.

2.3.2 Fingerprint-based fold alignments

During the processing of the Catalytic Site Atlas derived data set (see sub-section 2.3.4) it became obvious that the Dali structural alignment algorithm is too slow for large scale data processing in reasonable time, which is illustrated by the analysis of the data set comprised of 15,000 ligand-sensing cores that were aligned all against all. The resulting 15,164 x 15,164 = 230 million pairwise alignments were distributed over the 64 dual core processors (AMD Opteron 2.2 GHz) of an in-house Linux cluster. Although all programs were optimized for high performance computing, the cluster was running at 100% of its capacity for about 90 days before the alignments were finished, resulting in an average runtime of 4.3 seconds per alignment and processor (for technical details, see Experimental sub-section 2.6.2). The total computer time scales with the square of the number of cores because of the all-against-all alignments. Additionally, it was found that only 22 % of the Dali alignments actually gave a positive result, which means they returned a RMSD and a Z-Score value. Thus, 78% of the computational time (roughly 70 of the 90 days) was required for computing the structural alignment of structures that are not similar. For even larger data sets the computer time needed will increase significantly as well as the absolute time spent on the alignment of non similar structures. For the whole PDB with roughly 60,000 structures, computation of the all-against-all alignments with Dali would take roughly 60,000 x 60,000 x 4.3 s = 15,480,000,000 s or 179,167 days on a single processor equalling 1,400 days on the 128 processor Linux cluster. Obviously, the scalability of the PSSC procedure using the Dali structural alignment algorithm is limited in that respect.

One promising method to overcome this scalability problem and solve the clustering problems described above is the development and use of protein fold fingerprints. A ‘fingerprint’ in chem- and bioinformatics refers to a vector, each component of which referes to a so-called ‘feature’, for example one defined property of the molecule described by the fingerprint. In structural fingerprints applied in cheminformatics, for instance, often each component represents one structural moiety and equals ‘1’ or ‘0’ if the molecule described does or does not embody this moiety.

For the PSSC, the idea was to create a fingerprint that describes the sub-fold, that is the spatial arrangement of the secondary structure elements of a ligand-sensing core in relation to each other. For this purpose, a simple distance-based fingerprint is used that is based on all c_α-c_α distances in the ligand-sensing core. The procedure for fingerprint generation is shown in Figure 41 and technical details are given in the Experimental sub-section 2.6.4. By analogy to the Dali program, the c_α trace and DSSP information about the secondary structure elements were employed to describe the sub-fold of the ligand-sensing cores. After the core coordinates and

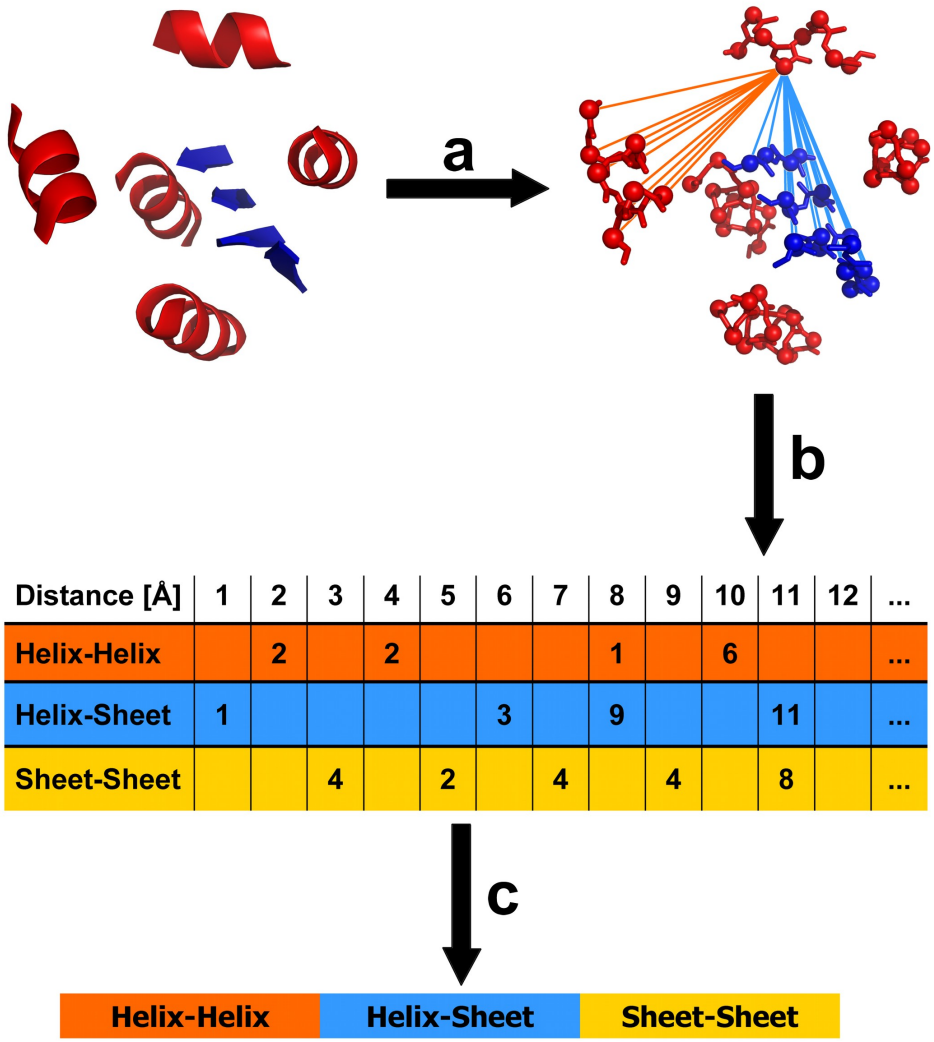


Figure 41: Generation of sub-fold fingerprints. The procedure starts from the structural information stored in the ligand-sensing core file annotated with the secondary structure elements predicted by the DSSP program. β-Sheets and α-helices are coloured in blue and red, respectively. a) All c_α-c_α distances are calculated and categorized according to the secondary structure elements they link. Not all distances are shown for clarity; only some distances for one c_α atom. Distances for the categories 'helix-helix' and 'helix-sheet' are marked in light red and light blue, respectively. b) These distances are then 'binned', that is they are assigned to equal distance ranges and the number of distances in each range is counted. Thus, three vectors are generated that contain the binned information for all categories of distances. c) In the final step, the three vectors are concatenated to yield the final fingerprint vector.

secondary structure elements have been read, the program calculates all c_α-c_α distances within the ligand-sensing core and categorizes them according to the secondary structure elements that both atoms belong to, for example as 'sheet-sheet', 'helix-sheet' or 'helix-helix' (see Figure 41a). In the second step, all distances of one category are binned, that is the distances falling into one particular range (= 'bin') are counted and the number is saved to the corresponding bin or, more precisely, to its position in the vector (see Figure 41b). This generates three vectors, one for each category of distances (helix-helix, helix-sheet, sheet-sheet) that are combined into

one fingerprint vector in the final step (see Figure 41c). Of course, other categories of c_α-c_α distances, for instance loop-sheet or loop-helix, could easily be integrated into the fingerprint as well. However, Dali exclusively relies on α-helices and β-sheets for its structural alignments and initial tests showed that these secondary structure elements provide a sufficient level of detail also in fold fingerprints. Several parameters influencing the binning had to be chosen: the minimum distance, maximum distance and the bin size. These criteria also determine the overall length of the fingerprint. The maximum distance allowed was determined empirically to 40 Å since no larger distances could be found in the ligand-sensing core set. As bin size, 0.5 Å were chosen based on the experience with structural distances and empirical testing. The value of 0.5 Å seems to be a good compromise between, on the one hand, the desire for the fine-grained detetction of differences between cores and, on the other hand, the problem of limited resolution in the structures. A minimum threshold for distances used in fingerprint generation was generated to reduce the noise introduced by the large number of small distances. The fingerprint describes the fold and the spatial arrangement of secondary structures in relation to each other. Therefore, the most significant distances are those between different secondary structure elements. Small distances mostly occur between residues in the same secondary structure element. In this case, these distances represent noise rather than valuable information. Following this line of argument, a minimum distance threshold of 5 Å was introduced. First empirical tests were promising to indicate that the discriminatory power of the fingerprint could be improved by this approach.

The fold fingerprint generation procedure was implemented in Java and used to generate sub-fold fingerprints of ligand-sensing cores. A second program for fingerprint comparison and compilation of similarity and dissimilarity matrices for clustering was also compiled in Java.

2.3.3 Addressing induced-fit in structures by molecular dynamics of ligand-sensing cores

Although PSSC has been successfully applied to homology models of proteins[43], structures experimentally determined by protein crystallography or nuclear magnetic resonance (NMR) form its basis. Whereas NMR structures do capture some dynamics of the protein structure incorporated in the multiple structure models generated from one measurement, protein crystal structures provide a freeze image of the continuous structural dynamics. Conformational changes occurring due to normal oscillation as well as induced-fit upon ligand binding may well change the sub-fold around the binding site significantly and, thereby, the PSSC cluster to which the protein should belong.

In collaboration with B. Charette (B.Sc.) and Prof. Dr. D. Berkowitz from the University of Lincoln, Nebraska, USA during their sabbatical in Dortmund in 2005/2006, the question was addressed whether molecular dynamics calculations of protein structures could provide a means to explore the conformational space for better classification by PSSC. The idea was to subject protein structures to molecular dynamics calculations and to cluster the resulting

ensembles of structures according to their similarity. Representatives from all the clusters would then be submitted to PSSC analysis and the resulting clusters would be studied.

As an example, the initial cluster derived by Koch *et al*.[43] was chosen that comprises of the dual-specificity phosphatase Cdc25A, acetylcholine esterase (AChE) and the 11β-hydroxysteroid dehydrogenases 1 and 2 (HSD1, HSD2). This cluster was well suited since it had been experimentally proven that similar compounds inhibit all cluster members. Moreover, none of the web-based databases for structural similarity, namely FSSP[192], CE[194] and VAST[215,216], was able to direcly retrieve AChE or HSD1/2 when using Cdc25A as a search template. Koch *et al*. had used a 'detour' *via* hydroxynitril lyase and the SCOP database to establish the cluster memberships. This may be due to the fact that the only available structure for Cdc25A is an apo structure (PDB code 1C25). Thus docking of a ligand structure and subsequent molecular dynamics simulations may be able to produce conformations of the ligand-bound protein. In light of the experimental proof for the PSSC cluster, one would expect these structures to exhibit a higher structural similarity to the other cluster members than the Cdc25A crystal structure.

Dysidiolide, a natural product inhibiting Cdc25A[229] was docked into the Cdc25A crystal structure using Autodock 3.0[230]. A subsequent molecular dynamics simulation with Gromacs[231] was perfomed, generating structures at 100 ps intervals. About 1,000 conformers were clustered with Gromacs for similarity. In the next step, the centres of these clusters served as search templates for the other members of the PSSC cluster. As opposed to the VAST search based on the Cdc25A crystal structure that retrieved only hydroxynitril lyase and methylene tetrahydromethanopterin dehydrogenase (MTHMP DH), the search based on conformer 668 retrieved all cluster members including AChE and HSD1/2. Moreover, conformers from other clusters also retrieved several cluster members and the hit frequency of the five proteins proposed by Koch *et al*. averaged over all conformer clusters was significantly high.

In a second PSSC analysis, Charette *et al*. started from cation-independent Mannose 6-phosphate/insulin-like growth factor II receptor (M6P-IGF2R)[232-234], a 300-kDa transmembrane receptor containing 15 homologous extracytoplasmic repeats. Its domains 3 and 9 are known to have a high affinity for M6P and bind M6P functionalized proteins. An initial VAST search starting from the three crystal structures available at that time, one apo structure (PDB code 1Q25) and two with bound M6P (PDB codes 1SZ0 and 1SY0), did yield some initial structural relatives. A subsequent molecular dynamics calculation of 1000 ps with conformation sampling every 100 ps yielded 10 conformers that were then subjected again to a VAST search. One conformer yielded the carbonic anhydrase (CA) from *N. gonorrhoeae* (1KOQ) as a potential PSSC cluster member. By analogy to the ligand-sensing core extraction described before, the conformations from the molecular dynamics simulation were clustered. From the centres of the 10 most populated structures, the ligand-sensing cores centred on M6P and with a radius of

25 Å were extracted and used as search templates in VAST searches. These searches yielded epidermal fatty acid binding protein (E-FABP) as an additional cluster member, a protein from the family of lipid binding proteins. It is associated with fatty acid signaling, cell growth, and cell differentiation. E-FABP overexpression has been observed in hyperproliferative skin diseases, such as psoriasis[235]. Additionally, it has been suggested as a cancer marker.[236,237] For more details of this work and experimental procedures, see the publication and Supporting Information by Charette *et al*.[238]

2.3.4 Clustering the Catalytic Site Atlas data set

To locate the active sites forming the centres of the ligand-sensing cores, the Catalytic Site Atlas[200,239,240] version 1.0 was used, a database of catalytic residues annotated from literature and PSI-Blast[241,242] sequence homology searches. After cross-checking structural information with the PDB, 25,509 catalytic sites were obtained belonging to 10,917 proteins that were then used in the ligand-sensing core extraction. 15 Å and 25 Å were set as the inner and outer core radius, respectively, yielding 45,371 ligand-sensing cores. This number is almost twice as high as the number of annotated catalytic sites due to NMR structures and homodimers, -trimers etc. For NMR structures, a ligand-sensing core was built for each structural model in the PDB file (see Experimental sub-section 2.6.1).

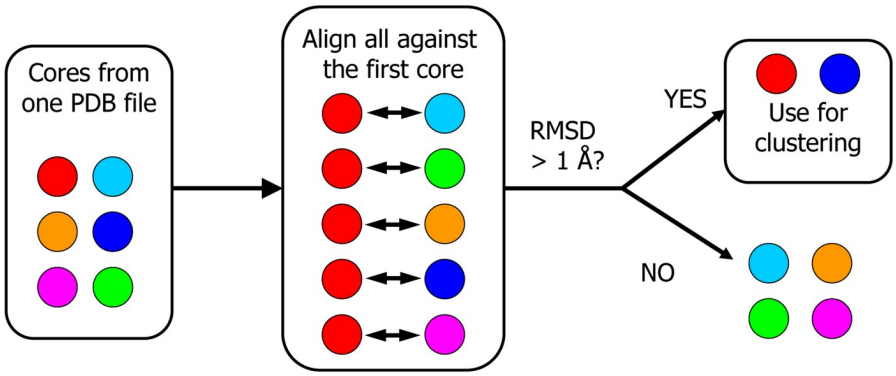


Figure 42: Pre-clustering procedure of the CSA data set. All cores relating to one PDB id, i.e., one protein structure, were taken and structurally aligned against the first core in the sequence using Dali. If the resulting RMSD value was below 1.0 Å, the cores were omitted. Otherwise they were kept together with the first core for further PSSC analysis.

Clearly, this set of ligand sensing cores was too big for running Dali structure comparisons all-against-all. It is well known, however, that the PDB contains many redundant structures[199], for instance homodimers, -trimers etc., multiple structures of the same protein, and NMR models that differ only very slightly. To remove much of this redundancy and reduce the number of ligand-sensing cores to a workable number, the ligand-sensing cores were clustered from each protein according to structural similarity. For this procedure, for each protein structure, i.e., PDB id, all related ligand-sensing cores were selected and aligned against the first core in the

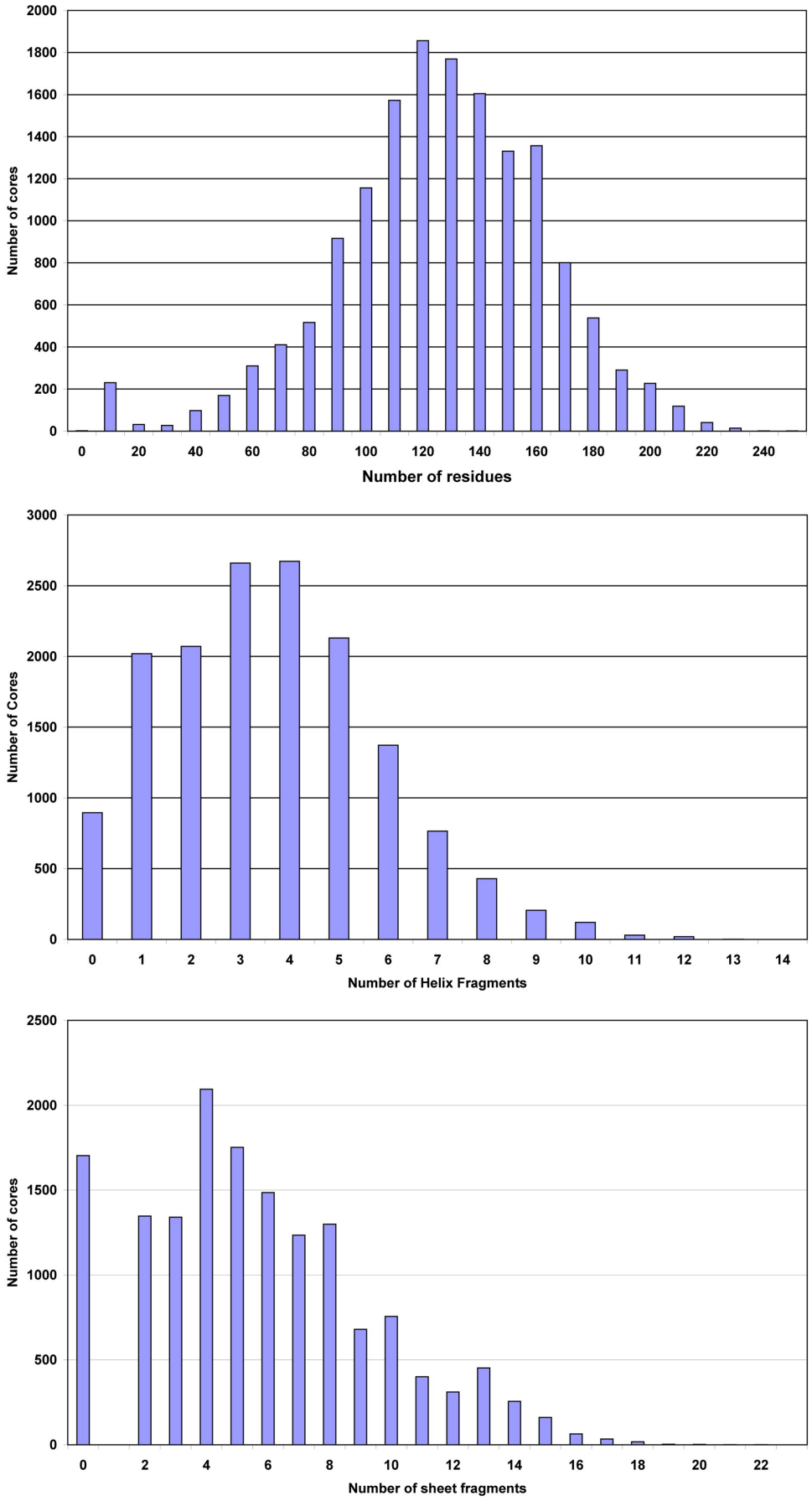


Figure 43: Distribution of ligand-sensing core properties. From top to bottom: number of residues per core, number of helix fragments per core, and number of sheet fragments per core.

sequence using Dali (see Figure 42). If the aligned cores yielded an RMSD of smaller than 1.0 Å, they were omitted. Otherwise they were subjected to further PSSC analysis together with the first core. This structure-based pre-clustering reduced the number of ligand-sensing cores from 45,371 to 15,164 without narrowing down the protein conformational space. The distributions of the number of residues per core, the number of helix fragments per core and the number of sheet fragments per core are shown in Figure 43. The core size distribution shows a maximum around 120 to 130 residues per core and tails off on both sides. The helix distribution has a maximum around 3-4 helices per core. There are also cores without helices as well as cores with up to 14 helix fragments. The sheet fragments are more evenly distributed and it is particularly noteworthy that no ligand-sensing core exists with only one sheet fragment whereas a sizeable fraction of cores does not incorporate any sheets at all. Combined, these distributions show that the ligand-sensing cores generated from the CSA are reasonable in size and do contain secondary structure elements; an important fact for the structure comparisons since both methods, Dali as well as the protein fingerprints are exclusively based on secondary structure elements.

The all-against-all structure comparisons were run as described in sub-section 2.3.1. In brief, the ligand-sensing core file set was distributed over the 16 nodes of an in-house Linux cluster. The alignments of $1/16^{th}$ of the cores against all other cores were started on each node and distributed internally over the eight processor cores. In about 90 days, each node calculated 14.4 million alignments generating 58 million text files containing the results. The total data volume far exceeded 1 TByte and could not be analyzed directly. Therefore, the important data were extracted and condensed into 64 files, a process that took several days on 16 processors. These files were then transferred to a MySQL database by the third program – a task that needed a further week on a single processor. The resulting MySQL database served as the data basis for the clustering in the next step.

As described before, the clustering algorithm was implemented as a Java program and directly queried the MySQL database via SQL commands. The fast retrieval of only selected alignments from the database reduced the computational time needed for the clustering by approximately one order of magnitude compared to clustering directly from the data files. This is mainly due to the fact that file input/output involving hard drives is significantly slower and the information is not indexed, i.e., all the information needs to be processed to find the desired bit. Processing the amounts of data generated by such approaches like large scale PSSC is close to impossible without high performance computers, for instance Linux clusters, and modern database systems. The clustering itself was performed within a few hours on a single processor.

As described above, the minimum Z-Score required was initially set to two as described by Holm and Sander.[196] However, based on the results of the first clustering where the minimal Z-Score value was set to two showed that a fairly large part of the alignments used for the

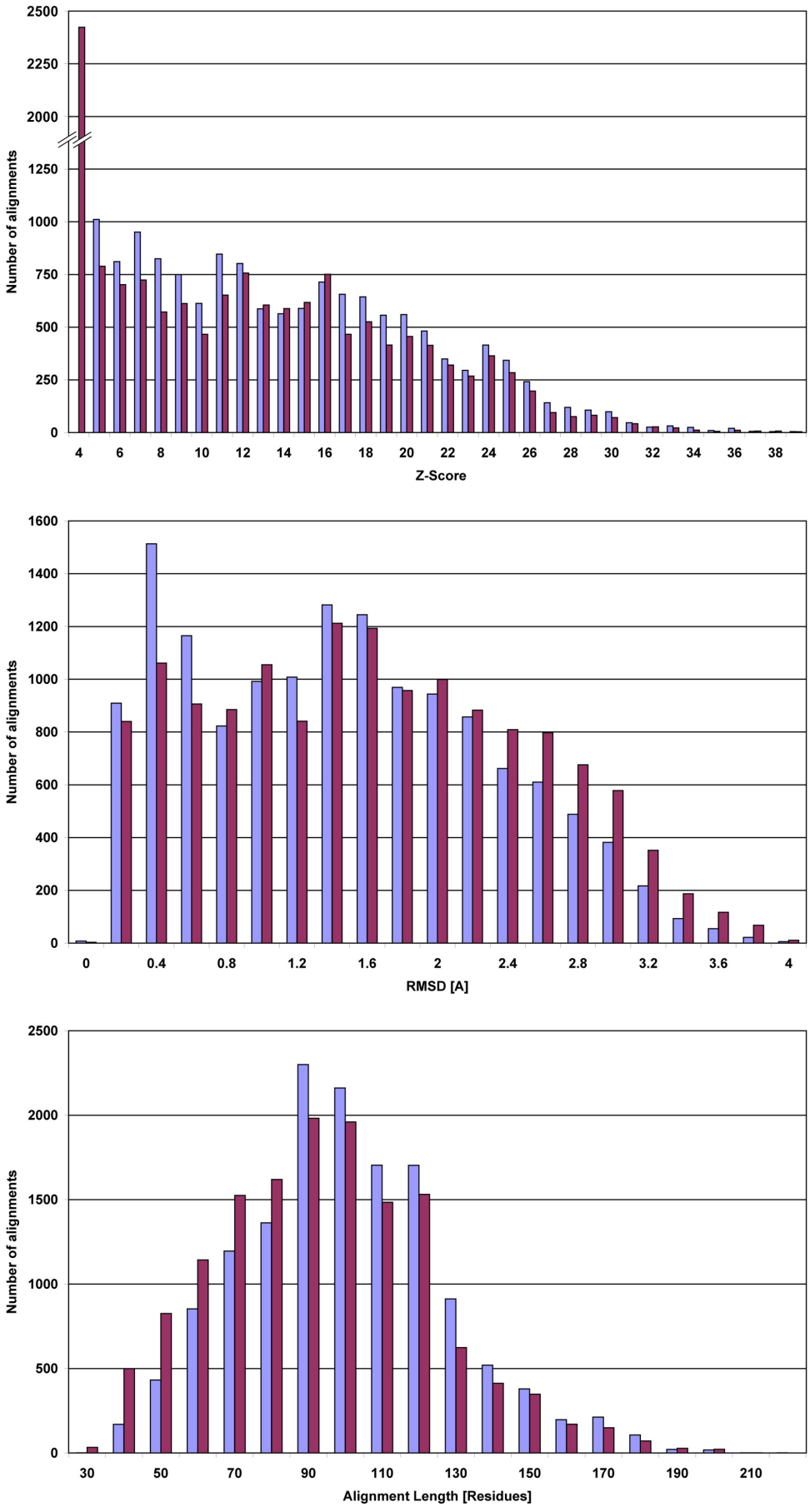


Figure 44: Distributions of Z-Score (top) and RMSD (middle) and alignment length (bottom) of all core-cluster centre alignments. Clustering was performed with a minimum required Z-Score of 2 (red) and 4 (blue).

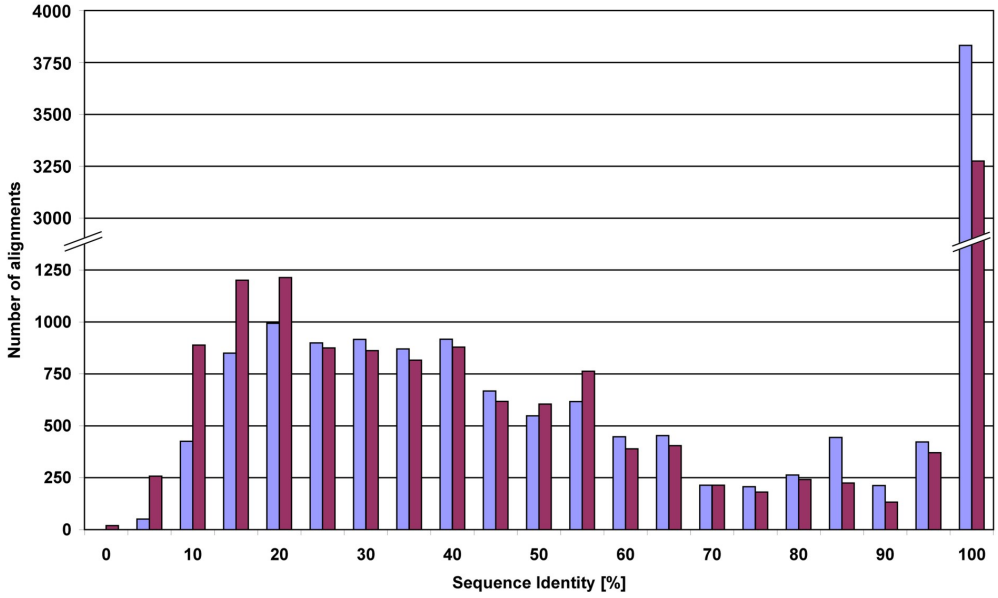


Figure 45: Distribution of the sequence identity of all core-cluster centre alignments. Clustering was performed with a minimum required Z-Score of 2 (red) and 4 (blue).

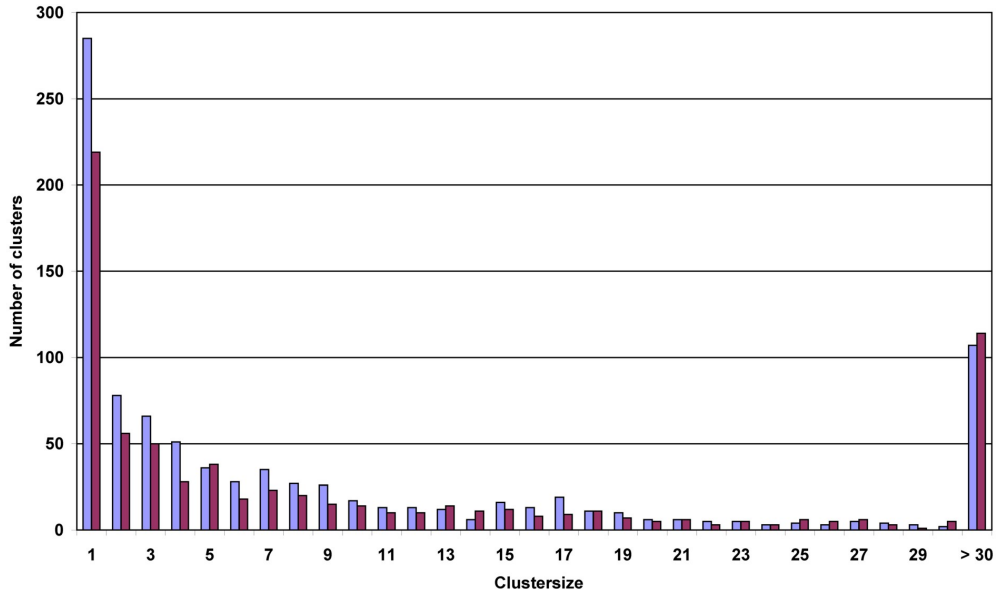


Figure 46: Distribution of results from the core-cluster centre alignments. From top to bottom: Z-Score, alignment length and sequence identity.

determination of cluster membership had Z-Scores between two and four (see Figure 44 top, red bars). Although not purely random according to the criteria of Holm and Sander the visual inspection showed that in most cases only a single secondary structure element of both cores is aligned, for instance one helix or one sheet. Setting the Z-Score threshold at four instead of two and re-running the clustering resulted in a different image (see Figure 44 top, blue bars). The minimum Z-Score for alignments assigning cores to their cluster is now larger than four but also the distribution shows a clear shift towards higher Z-Score values, in particular between 5 to 11,

which indicates a higher significance of the structural alignments. The RMSD values show a pronounced shift towards smaller values (see Figure 44 middle), especially to the very small values below one describing high structural similarity. The alignment lengths also shift towards higher values between 80 and 120 residues supporting the trend towards higher quality alignments (see Figure 44 bottom). One explanation for these trends could be discerned from the sequence identity distribution (see Figure 45). The increase of the Z-Score minimum threshold from 2 to 4 also led to significantly more co-clustering of cores with a high sequence similarity, particularly in the range between 95 and 100% sequence similarity. This is a strong argument in favour of the increase of the minimum Z-Score threshold since cores with a sequence similarity > 95% are highly homologous. Therefore, they are very likely to have highly homologous structures[243] and are expected to cluster together. As expected, the number of clusters increased by 180 from 734 to 914, while the number of singletons increased by 66 from 218 to 284. In general, the higher minimum Z-Score slightly decreases cluster sizes. Overall, these results indicate that the increase of the minimum Z-Score from two to four improved the overall significance and quality of the structural alignments that determine the cluster membership. Thus, the clustering process yields more sensible clusters.

The clustering results with a minimum Z-Score value of four looks very reasonable. The low number of 914 clusters out of the 15,164 ligand-sensing cores probably resembles to some extent the high redundancy in the PDB; a fact further reinforced by more than 3750 cluster assignments exhibiting a sequence similarity of > 95%. Although, the size distribution (see Figure 46, blue bars) indicates that 84% of the clusters contain less than 20 members including 284 (29.8%) singletons, 38 and 10 clusters have more than 100 and 200 members, respectively. These large clusters include a lysozyme cluster with 245 structures, pancreatic ribonuclease and related enzymes with 185 structures, glutathione S transferase with 99 structures, and many more clusters almost exclusively comprised of structures of the same protein.

To identify promising PSSC clusters, that is clusters whose member proteins are not predicted by known databases of structural similarity like SCOP[132,153,154] or CATH[244], a cross-validation with SCOP was performed. SCOP was chosen because it is manually curated and forms the gold standard of structure similarity and fold assignment.[245-247] The SCOP database provides a hierarchical genealogy to classify proteins according to the levels (from top to bottom) of class, fold, superfamily, family, and protein. Assignment of a protein to a branch of this classification is achieved by automatic structure comparison and high quality manual curation. For the cross-validation experiments, SCOP version 1.71 was used containing 27,599 protein structures classified into 971 folds, 1589 superfamilies, and 3004 families. The cross-validation was performed by mapping the PDB identifiers of the ligand sensing cores onto their corresponding SCOP classification using PipelinePilot (for technical details, refer to the Experimental sub-

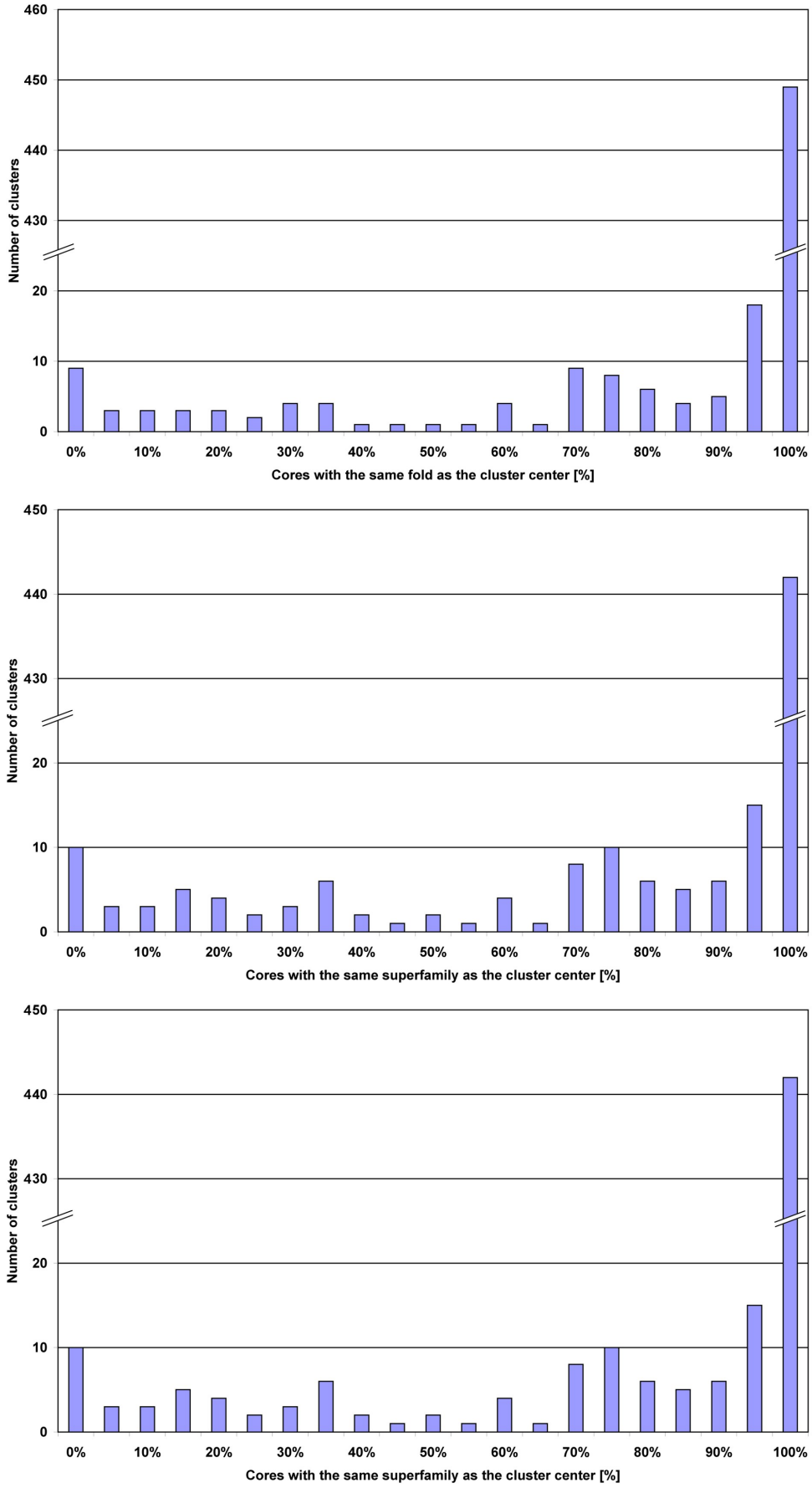


Figure 47: Statistics for cross-validation with the SCOP database. The diagrams show the fractions of clusters where a given percentage of members shares the corresponding SCOP annotation with the cluster centre. The annotations are (from top to bottom): fold, superfamily, and family.

section 2.6.5). SCOP data were available for 540 out of the 914 clusters (59%) and 12,560 out of 15,164 ligand-sensing cores (83%). Cores without annotation in the SCOP database were omitted in the cross-validation. As already expected from the structural alignment data discussed above, many clusters exclusively contain proteins that belong to the same SCOP classification. More than 440 clusters (= 81%) exhibit an average similarity in their SCOP annotation on the fold, family, and superfamily level of 95 - 100% (see Figure 47). Notwithstanding, there are 34 PSSC clusters (6%) in which less than 50% of the members share the SCOP classification with the cluster centre (for a complete list, see Attachment 7). These clusters or parts of them would be missed, therefore, by SCOP guided similarity analysis. They represent opportunities discovered by PSSC that might be exploited in the quest for new small molecule modulators of protein function.

2.3.5 Experimental validation of a cluster from the CSA set

From the number of clusters with heterogeneous SCOP annotations, one promising example was selected for experimental validation. The cluster comprised of 171 ligand-sensing cores,

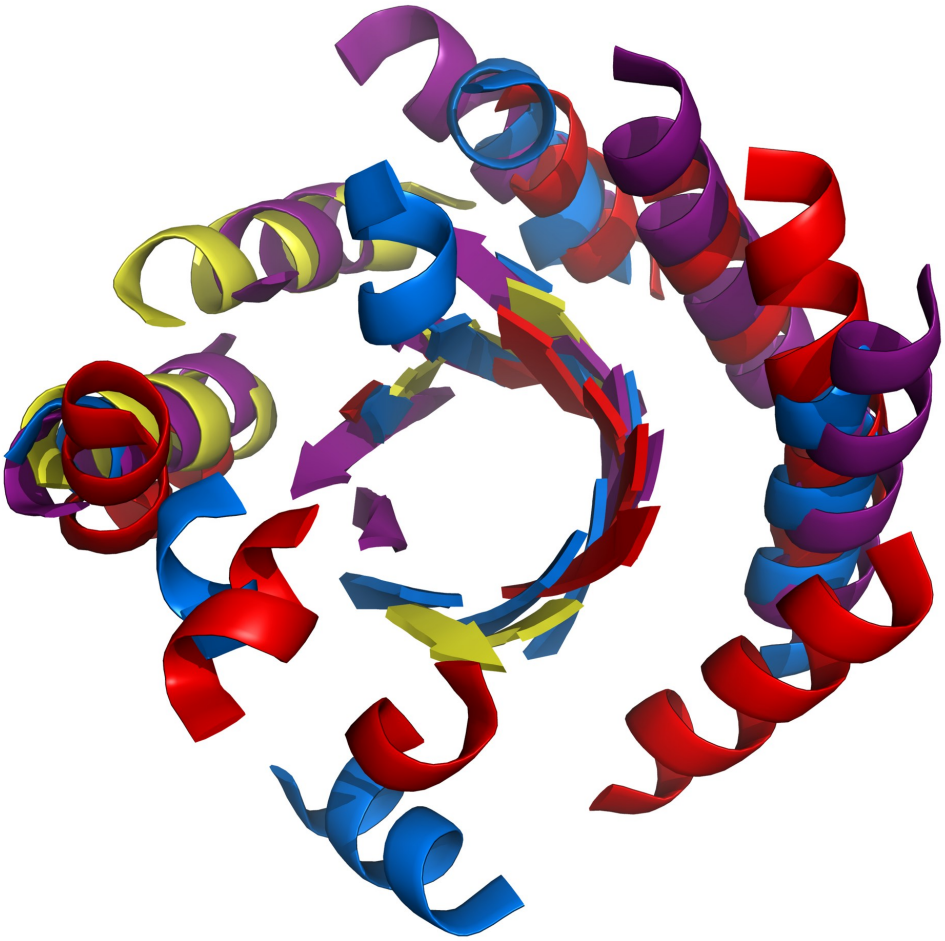


Figure 48: Aligned structures of four proteins from a promising cluster: pyruvate kinase (blue), dihydropteroate synthetase (red), xylanase (yellow), and methylenetetrahydrofolate reductase (purple). All structures are aligned with the cluster centre, another pyruvate kinase structure.

most of which are, like the cluster centre, structures of pyruvate kinase (PK), a metabolic enzyme catalyzing the reduction of pyruvate to lactate.[248,249] PK has been linked to metabolic diseases and cancer.[250,251] Interestingly, the pyruvate kinase clusters together with three different enzymes: dihydropteroate synthetase (DHPS), xylanase and methylenetetrahydrofolate reductase (MTHFR) (see Figure 48). DHPS catalyzes the reaction of 4-aminobenzoic acid and dihydropteridine-diphosphate to dihydropteroic acid, a precursor of tetrahydrofolic acid.[252,253] The folate pathway is an important target in anti-infective therapy and DHPS modulation is actively pursued in anti-infective drug development.[254,255] In particular, growing resistance in malaria has been linked to DHPS mutations and inhibition of DHPS is under investigation as anti-malaria therapy.[256,257] The third enzyme, xylanase, cleaves the linear polysaccharide β-1,4-xylan into its monomer xylose.[258-260] Xylanase is applied in the paper industry in the chlorine-free bleaching of wood pulp during the paper making process.[261,262] MTHFR is a flavoprotein belonging to the folate pathway. It reduces *5,10*-methylenetetrahydrofolate to *5*-methyltetrahydrofolate under the consumption of NADPH and is the only enzyme known so far producing *5*-methyltetrahydrofolate. [263-266] MTHFR has been linked to cardiovascular disease, schizophrenia, and cancer.[267-270] Current clinical research is centred on MTHFR inhibition as potential anti-cancer therapy.

Protein	PDB code	RMSD [Å]	Z-Score	Alignment length	Sequence Identity [%]
PK	1a5u	0.1	31.6	163	100
DHPS	1ad1	2.6	7.8	110	14
xylanase	1b30	3.3	4.2	83	12
MTHFR	1b5t	2.8	5.3	98	11

Table 10: Structural alignment data for the alignments of the cluster members with the cluster centre, which is another pyruvate kinase structure (PDB code 1a49). The higher Z-Score indicates a good alignment between pyruvate kinase and DHPS whereas the alignment with xylanase is less significant.

The structural similarity between the ligand-sensing cores of the cluster members described above is clearly visible in Figure 48. In particular, the central β-sheets align well in all cases. Of the surrounding helices only one is present in all cluster members whereas the others are partially missing. The statistical parameters of the structural alignment (see Table 10) indicate that PK aligns extremely well with the cluster centre because it is the same enzyme. DHPS also aligns reasonably well as documented by an RMSD of 2.6 Å and a Z-Score of 7.8 despite a very low sequence similarity of only 14%. The structural similarity of the cluster centre with xylanase is less pronounced, the Z-Score of 4.2 is barely over the minimum requirement value of 4. This is also visible in the structure overlay (see Figure 48, yellow structure) where several missing sheets in the central region can be identified. MTHFR exhibits a higher structural similarity than

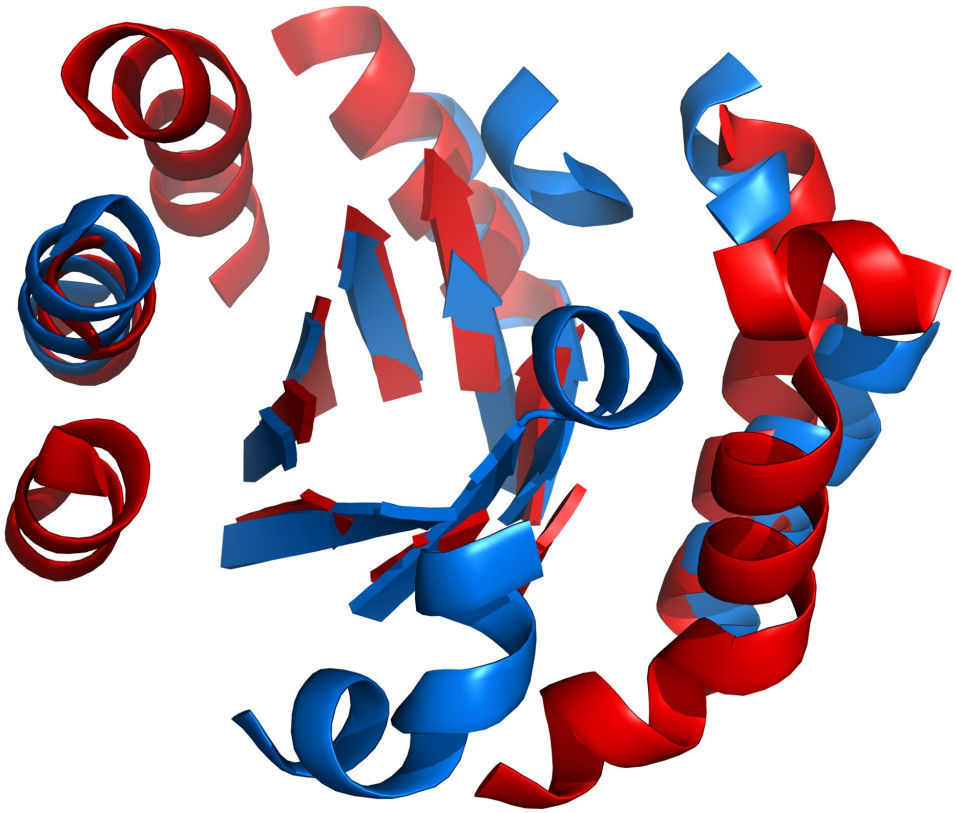


Figure 49: Structural overlay of the ligand-sensing cores of PK and DHPS. The well aligned beta barrel is clearly visible as well as three aligned helices.

xylanase but the sheets do also not align well. Therefore, experimental validation focused on the two most similar ligand-sensing cores, the PK and DHPS. For clarity, the overlay of both cores is shown in Figure 49 so that the good quality of the structural alignment becomes apparent.

PSSC suggests that the similarity between the ligand-sensing cores of PK and DHPS may lead to the binding of structurally similar ligands. Therefore, potential DHPS ligands and their scaffolds were tested as PK inhibitors. DHPS has been known to be inhibited by sulfones, sulfonamides[271] and sulfanilamides.[272] Many of these compounds were tested *in vitro* and also *in vivo* for their effect on *plasmodium falciparum*, the malaria causing parasite.[273] A number of SAR studies led to the development of potent inhibitors with documented *in vivo* activity.[274-279] Based on these known structures and the PSSC working hypothesis, a 740-membered compound library was purchased based on the sulfanilamide scaffold shown in Figure 50. The compounds were screened for inhibitory activity against PK from two different organisms, that is, from *bacillus stearthermophilus* and from *rabbit muscle*. Both initial screens were conducted at a fixed compound concentration of 100 µM.

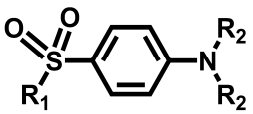


Figure 50: Sulphanilamide scaffold that served as structural template for the compound library to be tested against PK.

These pre-screens yielded 10 compounds with less than 50% residual activity that were subjected to concentration dependent measurements. However, all of these compounds showed IC_{50} values significantly larger than 10 µM. For the details of the screens, please see Experimental sub-section 2.6.7.

2.4 Discussion

2.4.1 The automated PSSC process: towards large scale database clustering

For the analysis of a large number of protein structures, a semi-automated PSSC process was successfully designed and implemented. The overall process consists of three functional modules: ligand-sensing core extraction, computation of structural alignment, and unsupervised clustering. Each module and its design and implementation, implications, as well as its scope will be discussed in the following paragraphs.

The design of the automated ligand-sensing core extraction process started with the development of criteria defining the cores in a systematic way. The mere extraction of all residues within a certain radius around the binding site did not suffice since it led to many incomplete secondary structure elements and fragmented strains all of which caused problems in the next step, the structural alignments. Simply increasing the core radius would not help because then parts of more distant secondary structure elements and linkers would be missed. To solve this problem, a four step extraction process was devised. In the first step, all residues within a given inner core radius were extracted and form the structural basis of the core that is selectively grown in the next steps to complete secondary structure elements and linkers. To limit this growth process, an outer core radius was introduced that is larger than the inner core radius. The space between the spheres defined by both radii denotes the “growth zone”. Only residues within this space can be added from the protein structure in the remaining steps. The second step completes the secondary structure elements by addition of their residues to the ligand-sensing core. Secondary structure elements are predicted by DSSP from the full protein structure to avoid that secondary structures are predicted inccorrectly due to fragmentation. As laid out above, secondary structure elements were only completed as long as their residues lay within the outer core radius. The same holds true for the chains linking fragments in the core that are added in the following step. In the fourth and last step, all fragments that comprise of less than five residues and do not form a secondary structure element are removed. Dali structure comparison is based on alignment of the secondary structure elements and, therefore, these small fragments are not needed but rather increase the total number of fragments in the core, which will interfere with the structural alignment.

Compared to a simple increase of the inner core radius and incorporation of more residues into the ligand-sensing core, the devised process generates ligand-sensing cores with fewer but well-defined secondary structure elements that are less fragmented. In the light of the PSSC

approach, these cores provide a better basis for the analysis since they are limited to the relevant but rather complete sub-fold around the binding site.

The structural alignments and similarity parameters were computed using the Dali software, a freely available program that has been found to yield the best structural alignments in independent tests.[222] The number of alignments that need to be calculated for an all-against-all comparison scales with the square of the number of cores. Although the individual pairwise structure comparison and alignment takes only a few seconds, the numbers of alignments easily add upt to hundreds of millions resulting in decades of computer time on a single processor. The method of choice to enable the computation of all the alignments within a reasonable time was the use of high performance computing, for instance heavy parallelization on a compute cluster. The most optimal parallelization technique is serial parallelization, i.e., the parallel execution of independent subjobs whose results are later combined. This technique does not require inter-process communication, rendering it rather fast and resource-saving. The millions of Dali alignments are an ideal case for the application of this parallelization method since each pairwise alignment is independent of all other alignments. Thus, a program was developed to distribute the basic data including the set of ligand-sensing cores and the programs over the cluster nodes, create a folder hierarchy on each node's local hard drive, and start the fraction of the total number of jobs that were assigned to this cluster node. After the calculations were finished, separately developed programs retracted, distilled and recombined the data and later stored it in a MySQL database. This 'brute-force' computational technique may not be very elegant but served its purpose by calculation of 230 million structural alignments within 90 days and could also cope with the resulting hundreds of GBytes of data. Nonetheless, the number of 15,164 cores clusters in this experiment was small compared to the overall number of protein structures in the PDB which amounts to about 58,000 structures as of June 2009.[280] Computing time for the Dali alignments of a set of 58,000 ligand-sensing cores would come close to three and a half years – running at full load of all processors 24 hours every day. Obviously, this is not practical. The results also showed that only a fraction of the structural alignments gave a successful alignment and similarity data. In the data set, this fraction was about 22%. If one only considers the significant structural alignments, that is, those with a Z-Score value of more than four, this number drops to 2.6%. This indicates that 97.4% of the computer time is spent on the computation of alignments of dissimilar structures. A fast but probably less accurate pre-filter could drastically reduce the number of alignments to calculate and, hence, the computer time needed. The protein fingerprints discussed in the next sub-section represent one possible approach to tackle this problem.

The combination of two values, RMSD and Z-Score, describing the quality and significance of the structural alignment proved to be a hurdle for the next step, the clustering. It was clear from the start that a manual 'clustering' by visual inspection of hit lists was out of question in this

work. Manual selection tends to be biased by the knowledge and preferences of the person making the selection but, even more important, it would take too much time. Manual analysis of the 15,164 hit lists resulting from the all-against-all Dali alignments was simply not possible. Therefore, unsupervised clustering was used. However, most clustering algorithms require a single distance measure that indicates the similarity or dissimilarity of two entities. Moreover, this distance measure needs to be continuous, which means that the measure must be applicable to all cases – from the very similar to the highly dissimilar. Both criteria are not met by the similarity measures RMSD and Z-Score. First, there are two values instead of one and combining them into one value proves difficult due to the different scales and trends. Moreover, both values carry a completetly different meaning and, hence, their combination would be 'cross-breeding apples with pears'. Second, neither the RMSD nor the Z-Score are continuous because they cannot be calculated for dissimilar structures. These facts ruled out the use of statistical clustering algorithms implemented, for example, in the free statistical software package R.[281] An adapted implementation of the OptiSim clustering algorithm developed by Clark[225] solved the problem and yielded sensible PSSC clusters after the optimization of the minimum Z-Score threshold. OptiSim is used to generate a diverse set of cluster centres and performs well in PSSC. However, it processes the cores sequentially and direct assignment of entities to one of cluster centres selected so far may not be optimal. As already laid out in sub-section 2.3.1, it may lead to the assignment of cores to clusters other than the one with the most similar cluster centre. In principle, this could be overcome by a re-calculation of cluster membership after the selection of all cluster centres although this has not been implemented yet.

A more promising way to overcome the problems associated with clustering and to facilitate the use of statistical clustering algorithms is the use protein fingerprints as described in the next sub-section. Protein fingerprints could in principle provide a fast and condensed description of the sub-fold contained in each ligand-sensing core. Moreover, many distance measures exist that are amenable to statistical clustering, which would grant access to a plethora of different, higly developed clustering algorithms.

Although the overal clustering process works well and facilitated the PSSC analysis of a set of 15,164 ligand-sensing cores, the in-depth analysis of interesting clusters needs to be done manually and include visual inspection of the structural alignments. The visual inspection was significantly improved by storage of the rotation-translation matrices calculated by Dali, which can be used to generate overlay images of many structures in a rapid automatic way. Cross-annotation of cores with protein details, SCOP categories and other information also improves the final analysis process that, nonetheless, remains time- and labour-intensive.

2.4.2 Protein fingerprints: scope and limitations

Protein fingerprints could address two shortcomings of the current PSSC process: the slow structural alignments and similarity calculations as well as the use of unsupervised statistical clustering methods. First, structural alignments can be significantly sped up by several orders of magnitude using structural fingerprints because the calculation of fingerprint similarity is much faster than computing 3-dimensional alignments. Second, almost all similarity measures for fingerprints that are currently in use are continuous and, hence, can be subjected to unsupervised statistical clustering. The overall applicability of PSSC would benefit greatly from the integration of such a method. As described in sub-section 2.3.2, structural fingerprints based on the c_α-c_α distances have been developed and implemented into a Java program. They were applied to the CSA core set and initial results look promising. Computing Dali alignments may still be necessary to calculate the RMSD and Z-Score values as well as to obtain the rotation-translation matrix needed for overlay images. However, clustering by fingerprint similarity only a fraction of alignments would need to be calculated compared to Dali-based clustering. To cluster N cores with Dali, N^2 alignments need to be calculated compared to only N alignments if the ligand-sensing cores are clustered by protein fingerprints. In the latter way, calculated RMSD and Z-Score values would serve as a built-in validation of the structural similarity determined by the fingerprint comparison.

The distance-based fingerprint that was designed and implemented in this work showed promising results in early comparisons of known PSSC clusters that compared well to the Dali method. Fingerprint generation by a Java program and subsequent similarity calculations were very quick compared to the Dali calculations and a speed-up by more than four orders of magnitude could be achieved.

There is some evidence that the simple distance-based fingerprints could be too simple to replace RMSD and Z-Score with similar accuracy. The fingerprint is based on the distribution of distances between c_α atom pairs. Therefore, if two cores are structurally similar, the distribution of these inter-atomic distances should also be similar. The converse argument may, in some cases, not hold true: a similar distribution could also belong to a ligand-sensing core where atoms occupy similar spatial locations randomly rather than because of structural similarity. This would then reflect the Dali alignments with a rather low Z-Score. One possible approach to overcome these problems is the development of more complex structural fingerprints involving more atoms and the spatial arrangement of them, for example a triangle of three atoms or a tetrahedron consisting of four atoms. These multi dimensional geometrical shapes would add the 2- or 3-dimensional arrangement of the c_α-c_α distances, which could improve the specificity of the fingerprint. Alternatively, one could try structural fingerprints for proteins that have already been implemented[185,282] although it is difficult to assess a priori if they are suitable for the problem at hand.

Neither a cross-validation with the results of the Dali method nor a fingerprint-based clustering were in the scope of this work. Development of more complex fingerprints was not possible within this work due to time constraints. These experiments should be undertaken in the future.

2.4.3 Clustering the Catalytic Site Atlas data set

One of the first challenges to meet was the definition of ligand binding sites in proteins. That was needed to locate and subsequently extract ligand-sensing cores from protein structures. At that time, the CSA offered an opportunity to get this information in a computer-processable form. After some additional data curation, for instance the removal of protein structures that were no longer accessible in the PDB, it served as the basis to generate a set 45,000 ligand-sensing cores that was narrowed down to 15,000 cores by a structure-base pre-clustering. PSSC analysis yielded about 940 clusters of differents sizes, many of which consist of mainly structures of the same protein reflecting the bias in both the CSA and the PDB. All clusters were cross-checked with the SCOP database to filter out clusters that could also be predicted using SCOP. This cross-validation showed that in the analysis, a larger fraction of the proteins defined as similar by SCOP also clustered together. Moreover, 34 clusters could be identified where SCOP predicts only 50% of the cluster members or less.

The clusters resulting from PSSC analysis of the CSA core set exhibit a large number of clusters comprised almost exclusively by structures from one protein. This directly reflects the accumulation of particular types of enzymes in the CSA. By its nature, the CSA is limited to enzymes since it stores catalytic residues. Although about 1,000 references are used for active site annotation, most of the catalytic residues were generated *via* sequence alignments using PSI-Blast searches. This clearly leads to an enrichment of structures of similar or identical proteins in the set. For a more representative set, the binding site annotation needs to cover more protein classes and structures. Relibase, a protein structure database marketed by the CCDC in the UK, identifies the small molecular ligands in PDB structures and annotates their binding sites.[283-285] Relibase offers a Python interface by which these binding sites can be extracted and stored. Initial work has been performed to compile a ligand-sensing core set based on all binding sites for ligands with a molecular weight between 100 and 1,000 Dalton. This set comprises about 100,000 ligand-sensing cores but was not analyzed further because it was too large to be clustered by the process established for the CSA set.

2.4.4 Pyruvate kinase and dihydropteroate synthetase - a PSSC cluster?

From the 34 clusters identified by the PSSC analysis of the CSA-based ligand-sensing core set, one sub-cluster was chosen comprising pyruvate kinase (PK), dihydropteroate synthetase (DHPS), xylanase and methylenetetrahydrofolate reductase. The analysis predicted the highest and most significant structural similarity for PK and DHPS. Hence these two enzymes were selected for the validation study. There are mainly three classes of DHPS inhibitors known in

literature: sulfones, sulfonamides and sulfanilamides. A compound library of 740 diverse sulfanilamides was acquired from a compound vendor and screened it against pyruvate kinases from two different organisms. The initial screen at a fixed compound concentration yielded 10 hits that were subjected to concentration-dependent measurements for validation. Unfortunately, none of the hits could be validated by the experimental setup in which the screen was run. This may be due to the relatively high concentration of ADP that had to be used in the assay to obtain a good signal (see Experimental sub-section 2.6.7 for details). Sulfanilamides have been described as ATP competitive inhibitors in kinases.[286,287] If they are indeed ATP (and ADP) competitive, a large excess of ADP may effectively deny the detection of compounds with fair potency. To test this hypothesis, other detection systems working with lower ADP concentrations were tried; the KinaseGLO® system from Promega and the homogeneous time-resolved fluorescence (HTRF) ADP Transcreener® from Cisbio. KinaseGlo® is based on a luciferase reaction that consumes the ATP in the assay solution and gives a corresponding chemoluminescent signal. HTRF Transcreener ADP is a displacement assay where a donor-labeled ADP is competitively displaced from an acceptor-labeled ADP antibody by ADP in the assay solution. The resulting reduction of fluorescence resonance energy transfer (FRET) is proportional to the amount of labeled ADP that has been displaced and, therefore, to the ADP concentration in solution. Both assays did not yield useable results for the detection of pyruvate kinase activity and inhibition in a concentration dependent manner.

One general problem could be that both assays are designed to detect normal kinase activity. Thus, they are most senstive for the detection of low ATP concentrations and high ADP concentrations since kinases catalyze the reaction from ATP to ADP. In contrast, pyruvate kinase catalyzes the transfer of a phosphate from phosphoenolpyruvate to ADP generating ATP. This results in high ATP and low ADP concentrations – the reverse situation from normal kinase assays. Both detection systems then work in the less sensitive ranges and close to their detection limit, which may explain why they did not work in this case.

Although it remains unclear whether the PSSC principle works for PK and DHPS in the case of the sufanilamides or not, in general, one cannot expect PSSC to work for all cases. On the one hand, it may well be that the structural similarity, albeit quite good for PK and DHPS, is not sufficient in this case. On the other, finding a suitable small molecule inhibitor often depends on the number of compounds analyzed since the probability of finding a hit is usually below 1%. Thus, screening a larger or different library of sulfanilamides may yield hits, the results of this campaign notwithstanding. It could also be that other compound classes, for instance sulfones or sulfonamides, would yield hits in this case.

Thus, the results presented before do not disprove the PSSC concept as a whole. They do, however, show the realistic sope of PSSC as a hypothesis generating tool that may work in

some but probably not in all cases. Moreover, it clearly demonstrates the challenges one might encounter when trying to prove or disprove a hypothesis experimentally.

2.5 Outlook

2.5.1 The future of PSSC

The semi-automated PSSC procedure designed and implemented in this work can be considered a major step forward towards the PSSC analysis of large data sets, possibly even the whole PDB. It established objective criteria and an automated process for the generation of ligand-sensing cores. The computation of structural alignments and similarity measures was parallelized and the final data transferred to a MySQL database. Finally, an automatic clustering method was implemented that provided unbiased and large scale clustering. The resulting clusters were then cross-annotated with their SCOP classification. Those clusters not predicted by SCOP were manually analyzed and experimental validation for one cluster was attempted.

Despite all these achievements, there is still substantial room for improvement. As discussed before, the CSA core set is biased towards particular enzyme families and, thereby, omits the structures of other proteins stored in the PDB. A PSSC analysis of the whole set of PDB structures would definitely be more interesting and challenging. The set of 100,000 ligand-binding sites extracted from Relibase could provide a viable basis for such an analysis although some quality control of the generated core structures would be required beforehand. Additionally, one could reduce the number of cores for the analysis by firstly pre-clustering all the cores with a very restrictive similarity criterion, thereby removing all structural redundancy. For this step information available from web-based databases for structural similarity, e.g., SCOP, could be used to define the clusters. Otherwise, one could also build a set of ligand-sensing cores largely devoid of redundance by starting from the ‘unique PDB’ based on sequence similarity. But as conformational changes might still occur even in structures with highly similar sequences, this set might miss some structural diversity present in the PDB files. Thus, a structure-based selection would be preferrable.

The structure aligment and similarity calculation process based on Dali works well but is definitely still too slow for processing large data sets. As mentioned above, a fingerprint based method has been developed that yielded some initial promising results. More validation studies and cross-checking with the Dali-based alignment data is needed to assess the accuracy of the fingerprints and the Tanimoto coefficient used as distance measure. Depending on these results it may be necessary to implement and use more advanced fingerprints emploing multiplets, for instance three or four c_α atoms as described by Abrahamian *et al*.[288] Such fingerprints may actually improve the accuracy of the fingerprint comparison, a prerequisite for meaningful clustering.

The subsequent clustering could be improved as well using statistical clustering methods as implemented for example in the open-source statistics software package R.[281] This depends, however, on the successful development of protein structure fingerprints and the identification of a suitable distance measure. Cross-validation of these clustering results with Dali similarity measures is mandatory.

Finally, the storage of all clustering results in a database and retrieval via dynamic web pages would facilitate PSSC analyses. In this way, the PSSC results could be made accessible to the broader scientific community, which might actually foster use of the hyptheses as well as experimental validation of the concept. Moreover, the data could be used to supplement other analyses and effectively be compared to other approaches sharpening its profile and its scope.

The experimental assessment of PSSC clusters and the hypotheses generated by the PSSC approach is a time- and resource-intensive undertaking that can hardly be pursued by one research group alone. To ‘prove’ or ‘disprove’ PSSC may both be impossible due to the large number of proteins, the relatively small fraction of those that have been structurally characterized and the number of positive or negative experiments needed to support either conclusion. It can be envisioned though that there is a more defined scope of PSSC, for example particular classes of proteins where PSSC works especially well or not at all. Reasons for such cases could be large conformational changes upon ligand-binding, induced-fit, highly dynamic protein structure among others. Exploring this scope would be highly interesting and increase the value of PSSC as a tool for small molecule ligand discovery.

The dynamic treatment of protein structures by molecular dynamics as described in sub-section 2.3.3 adds another interesting perspective to the future of PSSC. In principle, modern high performance computing allows dynamic modelling of a larger number of protein structures and subsequent clustering of the resulting conformations according to their structural similarity. Subsequent ligand-sensing core extraction and PSSC analysis may lead to novel clusters and interesting insights into the dynamics of structural relationships between proteins. Automated homology modelling[289-293] and protein structure prediction[294-296] are fields that progress rapidly towards more reliable and more accurate predictions. Since PSSC is based on the sub-fold rather than on the individual atomic positions of residues in the active site it may tolerate less accurate structures better than other approaches. Hence, one could even imagine integrating a repository of protein structure models like the Swiss-Model repository[291] with the PDB for an even more comprehensive analysis of protein structure space. The availability of faster high performance computers, for instance Linux clusters, will enable such analysis provided that faster methods for structural comparison and similarity based clustering become available.

2.6 Experimental

2.6.1 Ligand-sensing core extraction

The ligand-sensing core process was implemented in a program written in Perl. Parts of the initial design were done in collaboration with Wei-Zhe Hong during a one month internship and Mahesh Kulharia, a fellow Ph.D. student, who partly worked on the project for two months to get an introduction to bioinformatics.

```
1gos,0,GLY,A,13,manual,
2bxr,0,GLY,B,22,manual,
2ejr,0,GLY,A,287,manual,
1n9e,0,HIS,A,530,manual,
```

Figure 51: active site information format for ligand-sensing core generation

The program is run and supplied with three required parameters: the name of the file containing the active site definitions, the inner core radius and the outer core radius. Optionally, one can activate a detailed logging function as well as the output of all the intermediate cores from each step of the ligand-sensing core extraction.

The active site definition file contains the information about the active site residues of the proteins. This is stored in a comma-separated list with six columns containing the PDB code, site number, amino acid three letter code, chain identifier, residue number and extraction method, as well as an optional comment in a 7th column (see Figure 51). The PDB code is used to retrieve the PDB file containing the protein structure named “pdb<PDB code>.ent”. The site number indexes multiple sites per structure, for instance in homodimers. The individual residue of the active site is defined by its three letter code, the chain identifier, and its position number in the PDB file. Information about the source and comments are not strictly required but optional.

The program itself works as described by the following pseudo code:

```
get input parameters
read active site definition file
group entries by PDB id and site index
for each site
   if the structure file is found in the local structure directory
   then
      if it is an NMR file with more than one model
      then process every model individually and generate a ligand-sensing
         core for each
      else process the protein structure

      if all active site residues are contained in the structure
      then

         ******** extract inner core ********
         create list of all inner core residues
         create list of all outer core residues that are not in the
         inner core
         add all inner core residues to the ligand-sensing core
```

```
        ******** complete secondary structure elements ********
        read secondary structure information from pre-computed DSSP
        file
        if fails then create DSSP file and read the information
        create a list of residues in secondary structure elements


        for each residue in the secondary structure element list check
           if residue is in outer core
           then add it to ligand-sensing core

        ******** complete linker chains ********
        identify chains linking the fragments in the ligand-sensing core

        for each linker chain check
           if all residues are within the outer core radius
           then add linker to ligand-sensing core

        ******** remove small fragments not forming SSEs ********
        identify all chains in the ligand-sensing core
        for each chain check
        if it is smaller than 5 amino acids
        then check
           if chain is not in a secondary structure element
           then remove chain from ligand-sensing core
        write ligand-sensing core structure file in the pdb format
next site
```

Figure 52: Peudocode of the ligand-sensing core extraction program

The ligand-sensing core is based on the centre of mass (COM) of the c_α atoms of the active site residues. Residues within the inner core or the outer core are selected based on the distance between their c_α atom and the COM. Upon addition of all inner core residues to the (yet empty) ligand-sensing core, the secondary structure elements extending into the 'growth zone' between inner core radius and outer core radius are completed by incorporation of their residues into the ligand-sensing core. In the next step, linkers whose residues are located within the outer core radius are complete by adding their missing residues to the ligand-sensing core as well. In the final step, chains that consist of less than five residues and are not part of a secondary structure element are removed in order to reduce the number of chain fragments in the ligand-sensing core. Finally, the coordinates of the residues comprising the core are retrieved from the PDB file and written to the core file which is also in the PDB format.

If the command line parameter for the output of the intermediate cores is given at the program start, a pdb file with the current ligand-sensing core will be written after each step of the

process. Similarly, the extended log files will contain detailed information about detected SSEs, completed SSEs, detected linkers, completed linkers, and removed chain fragments.

2.6.2 Large scale structural alignments with Dali

The large scale calculations of the structural alignments were run via the queuing system Sun Grid Engine (SGE). The jobs were submitted to the queuing system via a perl script that evenly distributed the total number of alignments over all the cluster nodes, i.e., the individual server computers comprising the cluster. All data were stored at the local harddrive in each cluster node to reduce network traffic to an absolute minimum. A second program written in Java was run on each node once the Dali alignments were finished. This program checked the presence of all all the output files for each alignment to determine whether it was run at all. The completion of the calculation could be checked by a file named “dali.lock”. If this file is present, the calculation is still running or has been aborted with an error since the file is created at the program start and removed after completing the program. In the next step, this program extracts the information about the alignment result including RMSD, Z-Score and others from the result files of each alignment and saves them in a single central file on the cluster node. In parallel, all unnecessary result files are removed from the directory of this particular pariwise comparison. After finishing all the pairwise alignments of all cores against one particular ligand-sensing core, these results are compressed into a single gzip-file and stored for later use. A third program transferred the data from these centralized result files to a MySQL database. In total, the Dali alignments generated about 2.3 billion files with a total size of more than 1 TByte. The resulting MySQL database is about 10 GBytes in size without the translation-rotation matrices. Handling these volumes of data is not possible without a professional database system like MySQL.

2.6.3 Dissimilarity based clustering based on RMSD and Z-Score

The clustering algorithm was implemented in Java. Data was directly taken from and written to the MySQL database using the standard java database connector (JDBC) and SQL commands. The course of the program is shown in the pseudo code below.

```
Retrieve ligand-sensing core list from database
   For each core do
      Retrieve list of all alignments against the already chosen cluster
         centres with a Z-Score > 4 and an RMSD < 4
      If no alignment is found
      Then
         form a new cluster with the core as centre
      Else
         rank alignments by ascending RSMD and decending Z-Score
         Assign core to cluster on first position of list
      Write core and corresponding cluster centre to database
   Next core
```

Figure 53: Peudocode of the ligand-sensing core clustering program

The program first retrieves a list of all ligand-sensing cores in the database that is processed sequentially during the clustering procedure. For each core, a list is retrieved containing all the structural alignments of this particular core with the already selected cluster centres and a Z-Score > 4 and an RMSD < 4. If no alignments can be found, the core becomes the centre of a new cluster. If alignments are retrieved, the list is ordered by ascending RMSD and descending Z-Score and the core is assigned to the cluster centre at position one of the list of retrieved alignments. This assignment is then written to a table in the database. Iteratively, the list of cores is processed according to this procedure. The program clusters about 5 cores per second when run on a single processor with a locally installed database server.

2.6.4 Protein structure fingerprints

The program for the calculation of protein structure fingerprints was written in Java. Its pseudo code representation can be found below:

```
check the core directory for ligand-sensing core files
retrieve list of all ligand-sensing core files

for each core file do
   read the ligand-sensing core file
   check if the corresponding DSSP exists
   then
      read the corresponding DSSP file
   else
      issue an error message

   for each residue in core do
      if it belongs to a SSE
      then
         add to fingerprint residue list
         add c alpha coordinates to coordinates list

   for each residue in fingerprint residue list do
      for each other residue in fingerprint list do
         caclulate distance between c-apha atoms
         if minimum distance < distance < maximum distance
         then
            assign to distance list depending on SSEs of residues 1 and 2
            put distance in the corresponding bin in the distance vector

   concatenate all distance vectors to the fingerprint vector
   write fingerprint vector to a comma-separated text file
```

Figure 54: Peudocode of the program for calculation of protein structure fingerprints

The program assembles a list of ligand-sensing core files in a directory specified by the user. This list is then processed sequentially. First, the program reads the core structure file and the corresponding DSSP file - if it exists, that is. Then a list of all core residues within a secondary structure element is compiled and the x, y and z coordinates of the c_α atoms of these residues are stored in a second list. In the next step, for each residue in the fingerprint list, the c_α - c_α interatomic distance are calculated and binned according to their size and the secondary structure elements that both atoms belong to. In the last step, all these binned distance vectors are combined into a single fingerprint vector which is written to a comma-separated text file for future use.

2.6.5 Preparation of the Catalytic Site Atlas core set

The Catalytic Site Atlas version used in this analysis contains 100,599 catalytic residues assigned to 26,730 sites of 11,907 proteins. By cross-checking with the PDB 25,509 catalytic sites belonging to 10,917 proteins could be verified, completed with lacking catalytic residue information and were used for ligand-sensing core generation. For comparison: the contemporary version of the Catalytic Site Atlas contains 217,428 analytic residues that form 82,723 catalytic sites of 23,266 proteins. That resembles about three times as many catalytic sites in about twice as many proteins compared to the version used here.
The set of 25,509 catalytic sites yielded 45,371 ligand-sensing cores that were structurally clustered to reduce redundancy. The reduced set contains 15,164 ligand-sensing cores.

2.6.6 Preparation of the Relibase core set

A set of ligand binding site residues was extracted from Relibase, a structure database containing annotated structural data from the PDB. The script for binding site extraction was compiled to retrieve only binding sites that are occupied by ligands with a molecular weight between 100 and 1,000 Dalton. This filter was introduced because Relibase also identifies very small molecules resulting from the crystallization protocol, for example acetate or counter ions, as ligands and annotates their ‘binding sites’ as well. These pseudo binding sites are effectively removed by the lower molecular weight limit whereas the binding sites of proteins and large peptides are removed by the upper limit. The code for the script is presented in Figure 55.

```
import reliscript as rs
import re

hetatm = re.compile("HETATM")
term = re.compile("TERM")
ligset = rs.set('ligand')
outfile = open('/clz/work/wetzel/relibase/relibase_sites.out', 'w')
bscset = rs.set('chain',[]);
mw_min = rs.numeric_search('mol_wt',min=100.0)
```

```
mw_max = rs.numeric_search('mol_wt',max=1000.0)
ligset(mw_min)
ligset(mw_max)
i = 0;
for lig in ligset:
   i = i + 1
   if (i > len(ligset)):
      break
   if i % 10 == 0:
      print "record %07d " %i,"of %07d" %len(ligset)
   bs = lig.binding_site
   if (len(bs.chains) > 0):
      for bsc in bs.chains:
         if (len(bsc.residues) > 0):
            outfile.write(str(lig.compound_name)+'\t'+str(lig.n_atom)+'\t'+
               str(lig.mol_wt)+'\t'+str(lig.n_unit)+'\t'+
               str(lig.nucleic_acid)+'\t'+str(lig.cofactor)+'\t'+
               str(lig.covalently_bound)+'\n')

            for res in bsc.residues:
               outfile.write(str(res)+'\n');
      print "******* NEW SITE **********"
outfile.close
```

Figure 55: Reliscript program for the extraction of ligand binding site residues from Relibase

The Python-based script extracts a set all ligands with a molecular weight between 100 and 1,000 Dalton. In the next step, it extracts the binding site annotated in Relibase for each of these ligands, i.e., the residues within a 7 Å radius of the ligand. If any protein residues were retrieved, the compound name, number of atoms, chain, and several other parameters are written to a comma-separated text file. In the last step, the residues forming the binding site are written to the file as well.

The extraction of all binding sites with ligands of a molecular weight between 100 and 1,000 Dalton resulted in a set of about 1.5 million residues describing 67,240 binding sites in 19,464 protein structures.

2.6.7 Experimental validation of the pyruvate kinase / dihydropteroate synthetase cluster

Compounds:

All compounds were acquired from Chem Div[78] as 10 mM stock solutions in DMSO.

Pyruvate kinase assay (enzyme from *bacillus stearothermophilus*):

The lactate dehydrogenase (LDH) coupled pyruvate kinase assay was set up according to protocols from PubChem[76,77] and Sigma.[80] Firstly, 1 µL of compound solution in DMSO was dissolved in 49 µL of a solution containing all components except the pyruvate kinase substrate

PEP. Of this mix, 4x7 µL were transferred to the small volume measurement plate. After incubation for 15 minutes at 30°C, 7 µL of PEP solution were added to each well and the plate was measured in a continuous kinetics mode. Final assay concentrations were 50 mM imidazole (pH 7.2), 0.6 mM NADH, 0.4 mM ADP, 0.14 nM R5P, 50 mM potassium chloride, 7mM magnesium sulphate, 0.01% Tween 20, 0.2 U LDH, 0.07 U pyruvate kinase, 0.05% BSA and 2 mM PEP.

Activity data were normalized to a negative control without PEP instead of an inhibitor control and to a positive control without compound. Compounds with less than 50% activity in the pyruvate kinase assay were confirmed in a concentration dependent measurement using 11 concentrations with a 2-fold dilution starting at 100 µM. The IC_{50} values derived from these experiments are reported together with their standard deviation for three independent measurements.

Pyruvate kinase assay (enzyme from rabbit muscle):

The lactate dehydrogenase (LDH) coupled pyruvate kinase assay was set up according to a protocol from Sigma.[80] Firstly, 1 µL of compound solution in DMSO was dissolved in 49 µL of a solution containing all components except the pyruvate kinase substrate PEP. Of this mix, 4x7 µL were transferred to the small volume measurement plate. After incubation for 15 minutes at 30°C, 7 µL of PEP solution were added to each well and the plate was measured in a continuous kinetics mode. Final assay concentrations were 38 mM potassium phosphate (pH 7.6), 0.6 mM NADH, 0.9 mM ADP, 0.12 µM Fructose-1,6-diphosphate, 7mM magnesium chloride, 0.01% Tween 20, 0.1 U LDH, 0.06 U pyruvate kinase, 0.05% BSA and 2 mM PEP.

Activity data were normalized to a negative control without PEP instead of an inhibitor control and to a positive control without compound. Compounds with less than 50% activity in the pyruvate kinase assay were confirmed in a concentration dependent measurement using 11 concentrations with a 2-fold dilution starting at 100 µM. The IC_{50} values derived from these experiments are reported together with their standard deviation for three independent measurements.

3 General conclusions and outlook

3.1 Scaffold Hunter – a perspective

This work underlines the value of the scaffold tree concept as a chemically intuitive means to chart and navigate chemical space. The newly developed set of rules derived from synthetic and medicinal chemistry knowledge and the biology-based tree generation offer different, yet complementary views on the biologically-annotated chemical space covered by a given compound collection. Both approaches were successfully applied to identify novel chemotypes of inhibitors for pharmaceutically relevant targets. The elegance of the scaffold tree approach lies in the simplicity and the chemically intuitive character of the method itself, which renders it understandable and accessible to the educated scientist.

The development of Scaffold Hunter grants the community of scientists working in chemistry, biology and anywhere in between these sciences easy access to the full potential of scaffold trees and their application to chemistry and biology. The growing number of publicly available sets of large scale structure-related biological data, e.g. in PubChem[130] or StARLite[131], is prone to make a large impact on biomedical research, especially in academia. Scaffold Hunter can support this development as a versatile tool enabling literally thousands of scientists to explore these databases and tap their full potential for their own research.

Scaffold Hunter as well as Scaffold Tree Generator are both publicly available free of charge to foster their application in day-to-day research. They are also ‘open-source’ meaning that the source code for both programs is available enabling everybody to modify and extend the programs according to their needs. It is hoped that this mechanism will support the future development of Scaffold Hunter and add further functionality that is demanded by the scientific community. Additionally, it may also lead to the development of Scaffold Hunter or programs derived from Scaffold Hunter into completely unexpected areas of research and applications.

3.2 Towards Protein Structure Similarity Clustering on proteome scale

It was shown that application of PSSC to larger sets of protein structures is possible once the active site annotation is available. A semi-automated process based on newly developed criteria facilitates ligand-sensing core extraction for all subsequent analyses. Structure comparison and similarity calculation using the available DaliLite program were implemented in an optimized high performance computing fashion. Subsequent clustering was performed via an adapted implementation of a clustering algorithm. This process was successfully applied to a medium sized set of ligand-sensing cores albeit with a considerable computational effort.

Cross-validation with the SCOP database identified a sizeable overlap between both classifications, thereby validating the newly devised process. An interesting number of clusters was identified though, were PSSC finds a singficiant degree of similarity between the ligand-

sensing cores of proteins but SCOP does not. Although experimental validation of one selected cluster failed, this does not invalidate the concept as a whole.

During the PSSC analysis the bottleneck of the newly devised process proved to be the structural alignments using Dali. Although the time needed for one individual alignment is on the order of seconds, the total time for the hundreds of millions of alignments demanded by the analysis amounts to decades of computer time. Such problems of scaling are often encountered in computer science and also in natural science since the advent of high throughput technologies. Development of one possible method to address this problem has been started and initial results were described. The development of protein structure fingerprints that provide fast access to structural alignments and fold comparison may not only benefit PSSC but also other projects, e.g. SCOP by increasing the speed of database searches. Therefore, this methodology would be interesting to pursue and develop further.

The future of PSSC will depend on the successful use of the concept for inhibitor design on a broader scale. Moreover, the definition of structural similarity that yields successful PSSC clusters still has to be developed and confirmed. It might be that besides the sub-fold similarity around the binding site, a similar spatial location of the catalytic residues or even a similar mechanism of action is required. Further research needs to be conducted in order to sharpen the profile and clarify the scope of PSSC as a method target clustering and compound library design.

3.3 Merging chemical and biological space

The scaffold tree and PSSC themselves are cheminformatics- and bioinformatics-based complementary approaches for the design of biologically prevalidated compound collections which aim to improve the probability for successful discovery of small molecule ligands and inhibitors. The combined use of these two approaches, however, offers a new route towards biologically relevant compound classes. The scaffold tree of biologically prevalidated compounds, for instance natural products, drugs or agrochemicals, could be explored with Scaffold Hunter to explore paths towards simpler structures of known inhibitors with similar bioactivity. This can be achieved by the identification of virtual scaffolds in close proximity to activity ‘hot spots’, i.e. scaffolds representing highly active compounds, in the scaffold tree. Thereby Scaffold Hunter provides a systematic way to explore structural simplifications in biologically prevalidated chemical space – a task that largely relied on the training and experience of seasoned medicinal chemists. In parallel, PSSC analysis based on the structure of the protein target of interest would provide a target cluster that can be explored in both directions: biology as well as chemistry. From a biology point of view, the PSSC cluster provides a hypothesis which other proteins might be modulated by the same compound class. This knowledge can be exploited in several ways: first, one can screen the compound library against all targets in order to discover novel inhibitors for the other proteins and thereby exploit the

biological potential of the compound collection more thoroughly. Second, the screening on the other clusters directly yields information about compound selectivity, an important information in drug discovery but also in biological research where compounds are applied to cells to study underlying biological phenomena. From a more chemistry-based view, inhibitor classes of the other cluster member proteins can yield additional biologically prevalidated starting points in chemical space for library design. These results can, of course, be fed back into Scaffold Hunter and may well serve as additional criterion to prioritize and select promising virtual scaffolds for compound library synthesis and future research.

4 Additional scientific projects

Within the doctoral work presented herein, a broad range of projects in chemistry and biology were supported through application of computational and statistical methods. Those projects that resulted in one or more scientific publications in peer reviewed journals are shortly desribed in this section and the resulting publications are cited.

4.1 NMR-restrained conformational analyses of peptidic macrocycles

Many biologically relevant natural products comprise of macrocycles.[57] The bioactivity is often defined by the conformation of the macrocycles, which, in turn, is influenced by the ring itself and the ring-based substituents. Therefore, determination of the 3-dimensional structure of bioactive macrocycles generates valuable information for the design of active analogues. Structure determination is mostly based on 2D NMR data that is used to delineate distances between individual atoms. Such distances are then incorporated into conformational analysis calculations as distance constraints.

NMR-restrained conformational analysis was performed for two projects, the synthesis and characterization of a collection of biphenomycin analogues[297] and a collection of stevastelin C3 analogues.[298] In both cases protocols for conformational analysis incorporating the distance constraints determined by NMR were developed and optimized. This also included the identification of wrongly assigned distances where constraints could not be satisified. The calculated ensembles of conformations were clustered and the structures analyzed with respect to the influence of substituents on the overall conformation of the molecules. Moreover, the match of the results with experimental results was checked, as well as the constraints. In the case of the stevastelin C3 analogues, series of virtual compounds were modelled without constraints including phosphorylated and non-phosphorylated species.[299] Subsequent docking experiments were carried out to generate a hypothesis for the link between conformation and biological activity.

4.2 Determining substrate specificity of phosphatases with micorarrays

Within this project determination of the substrate specificity of phosphatases by microarray analysis was investigated. Short phosphorylated peptide sequences were immobilized on a microarray. After addition of a phosphatase in solution, the amount of remaining phosphorylated peptide was measured, thereby assessing the de-phosphorylation of the individual sequences. This project was supported during data evaluation as well as interpretation of the observed results using molecular docking calculations. The possible binding poses generated by the molecular docking program were clustered and analyzed for structural factors differentiating good from bad substrates. Moreover, concerns about the linker lengths as well as capping of the peptides were addressed and attenuated by docking simulations.[300]

4.3 Development of geranylgeranyl transferase II inhibitors

The protein geranylgeranyl transferase II (Rab geranylgeranyl transferase, GGTase II) transfers a prenyl to small GTPases from the Ras superfamily, a prerequisite for biological activity of the GTPases. GGTase II is considered a drug target in anti-cancer therapy. It is related but not structurally similar to the geranylgeranyl transerase I (GGTase I) and its structural relative, farnesyltransferase (Ftase). Several inhibitors of farnesyltransferase in late stages of drug development were shown to also inhibit GGTase II, which may actually be responsible for the clinical effect.[301]

In the early phases, the results from a first screen of several thousand compounds were re-evaluated and checked for their overall quality. Based upon the quality assessment, the screening results were discarded and the assay protocol was optimized and adapted to a high throughput screening robot in iterative rounds of testing and evaluation. A semi-automatic evaluation process for fixed-concetration screens and concentration-dependant measurements was designed, implemented and used for the evaluation of all screening data related to this project.[301,302]

Explanations for the observed bioactivity as well as selectivity patterns were sought using molecular docking and modelling techniques. The determination of the structures of co-crystals of GGTase II and peptide-based inhibitors was also supported by cross-correlation of putative poses with experimentally determined structure-activity relationships, as well as by molecular modelling. Delination of structure-activity relationships from co-crystal structures led to the development of novel, more potent peptidic inhibitors.[301,302]

From known crystal structures, partially with bound ligands, a hypothesis for the development of potent, selective small molecule inhibitors of GGTase II was developed. This hypothesis was refined and confirmed by the generation and molecular docking of virtual compound libraries. In collaboration with Dr. Robin Bon, synthetically feasible structures that scored well in the docking experiments were selected for future synthesis and biochemical evaluation. This project is still in progress and pursued by Anouk Stigter (M.Sc.) and Dr. Robin Bon.

4.4 PSSC cluster with APT1: from PSSC to chemical biology

To experimentally investigate a PSSC cluster identified by M. Koch[303], a library of β-lactones derived from tetrahydrolipstatine, a marketed drug inhibiting gastric lipase, was synthesized by Dr. Frank Dekker. The design of this library was supported by iterative cycles of delineation of structure activity relationships and structure based compound design. The best structures were varied in virtual libraries, evaluated by molecular docking and results were cross-corelated with bioactivities determined by biochemical assays. Several additional analyses were carried out to verify and precisely determine the structural similarity between the ligand-sensing cores of acyl protein thioesterase 1 (APT1) and gastric lipase and, hence, their membership of the same cluster.

So far, several covalent inhibitors of APT1 resulted from this project, which were characterized with respect to their binding mode and kinetics. These analyses were supported by iterative circles of experimental design and statistical evaluation of the results.

The β-lacton based inhibitors were then successfully used for the illucidation of the biological role of APT1 in cells in a chemical genomics approach. It was proven that APT1 plays a key role in the depalmytoylation of Ras and, thereby, in modulation of its activity. The findings identify APT1 as an emerging drug target in the treatment of cancers, in which Ras is constitutively activated.[304]

5 Summary

Chemical genomics, i.e. the use of small molecule modulators of protein function to study the underlying biological processes, lies at the heart of chemical biology. Therefore, the development of small molecule libraries enriched in biological relevance is a key prerequisite for chemical biology research. Natural products can be seen as a group of compounds selected by evolution for their ability to bind to various proteins and, therefore, as biologically prevalidated. The work presented herein aimed at the development and application of computational approaches for compound library design. It included the build-up of the necessary computational infrastructure, method development, and experimental validation. The methods that were developed build on the structural complementarity of small molecule and protein space to map and explore biologically relevant parts of chemical space as well as the corresponding protein targets.

5.1 Cartography of and Navigation in Chemical Space

This section will describe development and application of methods based on the chemical space concept, i.e. the analysis and exploration of large collections of structural data.

5.1.1 Scaffold Tree

The “Structural Classification of Natural Products”[23] was developed in collaboration with Novartis to gain an overview over the structural templates most abundant in natural products. The classification is based on chemically meaningful scaffolds, i.e. all rings and connecting linker chains, as well as all double bonds directly attached to these. Iterative deconstruction by one ring at a time guided by a set of rules generates a branch of scaffolds rooted in the one ring scaffold. In this hierarchy, the smaller scaffold is termed the “child” scaffold and the larger scaffold the “parent”. Processing of the Dictionary of Natural Products[5] produced a tree-like diagram depicting natural product structural space. Each structure is represented by its molecular scaffold as described above. Scaffolds populating the same branch may also share other properties, i.e. bioactivity.[89] The term “brachiation” describes movement from larger scaffolds towards smaller scaffolds while keeping similar biochemical activity.

A second, more generic set of rules based on medicinal and organic chemistry knowledge was developed in collaboration with Novartis to construct more generally applicable scaffold trees.[56] Scaffolds that do not represent molecules in the dataset are incorporated into the scaffold tree and termed “virtual scaffolds”. Moreover, the new set of rules renders the generated scaffold trees independent of the data set, it facilitates the merging of scaffold trees generated from different sets of compounds.

5.1.2 Scaffold Hunter

Although scaffold trees chart chemical space in a chemically meaningful and intuitive way, their application to inspire and direct synthetic organic chemistry was limited by the static nature of the tree image. This image can show only limited information and does not allow for annotation with other properties, i.e. biochemical activity. Therefore, a joint student project with the Chair of Algorithm Engineering at the Technical University of Dortmund was designed and initiated in order to develop an interactive scaffold tree browsing program named “Scaffold Hunter”. Project design, as well as supervision and guidance of the project group of twelve computer science students through the one year project were part of this work.

Scaffold Hunter automatically generates a visual representation of the scaffold tree based on the data read from a database. It enables navigation in the scaffold tree including filtering, zooming, colour shading according to properties and bookmarking. A guided tour to scaffold hunter is offered in the Supporting Movie 2 supplied with the publication in Nature Chemical Biology.[123] The scaffold tree database can be easily generated with a second tool, the “Scaffold Tree Generator” written by S. Renner from any SD file, a standard open file format for molecular structures. Scaffold Tree Generator was modified within this work, in particular errors were fixed and the database connection was updated.

5.1.3 Finding and filling gaps in chemical space

As described above, the new rule set also allowed for virtual scaffolds that do not represent compounds in the data set. These scaffolds complete the tree and, hence, represent gaps in the chemical space of the analyzed data set. Such gaps can be exploited by identification of promising virtual scaffolds by their proximity to scaffolds representing potent compounds. A first proof-of-concept was obtained by a retrospective study extracting promising virtual scaffolds from PubChem data and searching for compounds incorporating these scaffolds that are described as active in Wombat, a database annotating small molecules with bioactivity data from literature. To obtain a prospective proof-of-concept the data set from a pyruvate kinase screen stored in PubChem was analyzed with Scaffold Hunter. Out of the 65 promising virtual scaffolds identified, small focused libraries based on four of these scaffolds were acquired and tested as modulators of pyruvate kinase activity. The biochemical screen yielded eight confirmed hits that had not been described as modulators of any protein before according to a SciFinder search. This demonstrates how Scaffold Hunter enables scientists to transform data into knowledge, and finally, new science. Scaffold Hunter, as well as the validation studies are published together with all the data.[123] The programs are available free of charge under an open source license via www.scaffoldhunter.com whereas the sample database and installation instructions can be downloaded from the Max Planck edoc server: http://edoc.mpg.de/display.epl?mode=doc&id=429252.

5.1.4 Exploring Natural Products: the γ-pyrones

One natural application of Scaffold Hunter enabled by the new set of rules is the comparison of different sets of compounds by merging of their scaffold trees. This may also be used to annotate one of the sets with the properties of the other, i.e. protein target information, which would be valuable for library design. In a first model case, the natural product chemical space was annotated with target information from the WOMBAT database. The branch of the γ-pyrones showed promising biochemical activities and was synthetically accessible. Moreover, testing compounds from different hierarchy levels of this branch may provide an example for brachiation in the *O*-heterocyclic sector already proven for the carbocyclic[23] and *N*-heterocyclic[44,48] parts of the scaffold tree. A library of higher γ-pyrones (2- to 4-ring scaffolds) was compiled and acquired, while a compound collection centred on the root scaffold of the γ-pyrones' branch was synthesized by Dipl.-Chem. Wolfram Wilk and Dr. Samy Chammaa. The retrospective analysis of the Wombat database had yielded some cases, where brachiation over three hierarchy levels was known. For prospective testing the monamine oxidases (MAO) subtype A and B, the signal transducers and activators of transcription (STAT) proteins and acid sphingomyelinase were selected as targets because activity was described for one or two scaffold types of the γ-pyrone branch. Subsequent biochemical testing confirmed active compounds for all proteins tested. The MAO A and B screens were carried out in-house together with Dipl.-Chem. Wolfram Wilk, whereas screens against the STATs and sphingomyelinase were performed by collaboration partners. All together, the screens yielded a significant number of selective hits, in particular 60 and 35 hits for MAO A and B, respectively, with IC_{50} values between several hundred nanomolar and 10 µM. The STAT screen identified two novel inhibitors that are similar in structure to known inhibitors but exhibit different selectivity patterns. One of the hits was shown to be active in a more demanding cellular assay. Three compounds with good inhibitory activity against acid sphingomyelinase and high selectivity for the acid over the neutral isoenzyme were identified. Additionally, these compounds possess much more favourable molecular properties than other known small molecule inhibitors. These results were published[126] and indicate that Scaffold Hunter may also be well suited to annotate compound sets with putative targets and thus focus screening and synthesis efforts. Possible target clusters for scaffold families can also be identified via the PSSC approach described in the next section.

5.1.5 Outlook

Future extensions of Scaffold Hunter are described, e.g. several mechanisms to enable the generation of scaffold trees according to customized set of rules, also invoking biology-guided scaffold trees as developed by S. Renner.[89] Additional improvements include a substructure search engine, advanced filtering capabilities, as well as a version facilitating interactive modification of rules with instantaneous generation and visualization of the resulting scaffold

tree. Application of Scaffold Hunter for the classification of other kinds of molecules, for instance catalysts, is discussed, as well as extension to other domains, e.g. protein structures. Finally, the possible extension of the scaffold tree for the generation of chemically meaningful natural product fragments is described. Such fragments possibly facilitate the exploration of nature's diversity with a reasonable synthetic effort.

5.2 Exploration of Proteomic Space – Protein Structure Similarity Clustering (PSSC)

This project aims at exploiting the structural complementarity between proteins and their small molecule modulators in order to map biologically relevant small molecule scaffolds onto clusters of potential target proteins.

5.2.1 State of Previous Research and Aims

Protein binding sites and their small molecule modulators need to be structurally complementary to each other. Hence, structurally similar binding sites should bind similar ligands which is well known, e.g. in kinases. However, similarity at the atomic level can be problematic due to conformational freedom, e.g. of protein side chains, or experimental limitations like crystal structure resolution. Therefore, similarity needs to be defined on a more abstract level; namely the scaffold. "Scaffold" in this context means a rather rigid 3-dimensional framework orientating the attached substituents that form most of the interactions. The scaffold is present in both, proteins and small molecules, represented by the backbone of the peptides spatially organized by protein folding and the often fused ring systems in small molecules, respectively. Since in both cases, the interacting substituents of the scaffold can be quite diverse despite high scaffold similarity, similarity leads to the starting point of a diverse compound collection to match the diversity of the interacting protein side chains.

Initially, the Protein Structure Similarity Clustering[43] approach used by Koch *et al* was based on manual fold comparison using online databases, like the FSSP[192,195] database, CATH[244] or SCOP.[132,153,154] Clusters were generated manually by overall fold similarity searches and subsequent visual inspection for structural similarity of the binding site. This made PSSC very laborious and user-dependent since many parameters, e.g. the size of the binding site, were not clearly defined.

The aim was to develop a more defined and automated process for PSSC and apply it to a larger set of protein structures, ideally the Protein Data Bank (PDB) as a whole. The clusters identified from such an approach should be validated experimentally to determine the scope of PSSC.

5.2.2 Method Development – Automated PSSC

The newly devised PSSC process centres on the structural alignment of "ligand-sensing cores", spherical cut-outs around the binding site, instead of full proteins. This drastically reduces false

positive results where similarity is present remotely of the binding site. Definition of criteria for the size of ligand-sensing cores was based on an evaluation of catalytic sites annotated in the Catalytic Site Atlas (CSA).[200] A self-written computer program automatically extracted ligand-sensing core structures optimized for subsequent structure comparison from PDB files using the binding site information from the CSA. Structural alignments were computed with DaliLite.[219,220] Results are comparable with those of earlier analyses since the FSSP database was also compiled with DaliLite. This alignment data is then clustered by a statistical algorithm implemented in Java. All programs used in the PSSC process were written within this work and optimized for high performance, except for the DaliLite.

Extraction of a set of ligand-sensing cores based on the CSA and pre-clustering according to structural similarity (RMSD < 1Å), reducing the redundancy found in the PDB, yielded 15,000 cores. Computation of the structural alignments (each core against each) took 2 full months on a 128 processor linux cluster. The results were stored in an SQL database for further analysis. Since the Dali algorithm calculates alignments only in successful cases and, hence, does not generate a complete similarity matrix, statistical clustering algorithms cannot be used on the data. Therefore, the data was clustered by adapted implementation of the OptiSim[225] clustering algorithm. Comparison with the SCOP database, the “gold standard” in structural similarity, yielded a large number of clusters that were also classified as similar by SCOP but also 33 clusters where SCOP predicts only less than half of the cluster members.

For experimental validation, two proteins from a PSSC cluster, pyruvate kinase (PK) and dihydropteroate synthetase (DHPS) were chosen and a library of 740 sulfanilamides, a known class of DHPS inhibitors, was screened for PK inhibition. Unfortunately, the screen did not yield any hit compounds, which may be due to the assay conditions of the screening system used but could also imply that PSSC does not work in this case. A final conclusion about the applicability of PSSC is not possible based on these results.

For large scale PSSC analyses, a fast structure comparison method based on protein fingerprints was developed. This method uses an abstract representation of the ligand-sensing cores as vectors that makes it several orders of magnitude faster than Dali. Cross-validation showed the new fingerprint based similarity in fair agreement with the Dali data for the 15,000-membered core set. Most likely, the chosen fingerprints were too simple and did not sufficiently represent the 3D arrangement of secondary structure elements. Implementation of a four-point fingerprint should address this issue and improve accuracy. Additionally, the similarity measure used with these fingerprints, the Tanimoto similarity index, can be computed for all alignments rendering the resulting similarity matrix amenable to statistical clustering with the free software R.

5.2.3 Method development – PSSC with dynamic protein structures

One inherent drawback of the PSSC method is the demand for a structure of the protein of interest. Although PSSC has been shown to work with homology models[43] induced fit might still be problematic. The aspect of integrating dynamics into PSSC by molecular dynamics calculations was investigated with B.D. Charette from Prof. Berkowitz‘s group in Lincoln, Nebraska during his research visit to the group. It was successfully shown that molecular dynamics could facilitate the transition from an apo protein to a ligand bound structure simulating the induced fit needed for subsequent PSSC analysis.[238]

5.2.4 Outlook

Development of more complex structural fingerprints based on the results discussed herein is described. Such a fingerprint-based technology enables the structural comparisons of a set of 67,000 ligand-sensing cores based on ligand binding sites extracted from the ReliBase.[284,285] The PSSC analysis of this set of binding sites representing the structurally known proteome will grant insights into structural relationships between proteins and enable the design of inhibitors targeting novel targets. Experimental validation of large scale PSSC results is needed to determine the scope of the concept. Extension of the method can be envisioned via integration of dynamic structures from molecular dynamics calculations, as well as automatically generated structural models for proteins where no experimentally determined structures are available.

5.3 Miscellaneous projects

During the doctoral work present herein, several other projects were supported as well. This work includes NMR-restrained conformational analysis of macrocycles[297,298] using force fields as well as docking of inhibitor structures to gain insights into experimentally observed selectivities.[300-302,305] Further work included the evaluation, statistical analysis and quality control of biochemical screening data and subsequent optimization of the experimental screening protocol.

6 Zusammenfassung

6.1 Einführung

Chemische Genomik, d.h. die Beeinflussung der Aktivität von Proteinen durch organische Moleküle zur Untersuchung grundlegender biologischer Vorgänge, ist eines der zentralen Forschungsgebiete der Chemischen Biologie. Die Entwicklung biologisch hoch relevanter Molekülbibliotheken stellt daher eine wesentliche Grundvoraussetzung für chemisch-biologische Forschung dar. Naturstoffe wurden im Laufe der Evolution auf Grund ihrer Fähigkeit, an verschiedene Proteine zu binden, selektiert und sind daher biologisch vorvalidiert. Die hier vorgestellte Arbeit zielte auf die Entwicklung und Anwendung computer-gestützter Verfahren zur Entwicklung von Verbindungsbibliotheken ab; einschließlich der Etablierung notwendiger IT Infrastruktur, der Methodenentwicklung sowie experimenteller Validierung. Die entwickelten Methoden basieren auf der Analyse komplementärer chemischer und biologischer Strukturräume zur Kartierung und Erforschung biologisch relevanter Teile des chemischen Strukturraumes sowie der dazugehörigen Zielproteine.

6.2 Kartographie und Navigation im chemischen Strukturraum

Dieser Abschnitt beschreibt die Entwicklung und Anwendung von Methoden, basierend auf dem Konzept des chemischen Strukturraumes, zur Analyse und Nutzung großer Strukturdatensätze.

6.2.1 Das Baumdiagramm der chemischen Gerüststrukturen

Die „Strukturelle Klassifikation der Naturstoffe“ (SCONP)[23] wurde in Zuammenarbeit mit Novartis entwickelt, um einen Überblick über die häufigsten Strukturtypen in Naturstoffen zu gewinnen. Die Klassifikation beruht auf chemisch sinnvollen Gerüststrukturen, d.h. den Ringsystemen, allen Strukturen, welche Ringe miteinander verbinden, und allen Doppelbindungen, die von den genannten Strukturen ausgehen. Die schrittweise Zerlegung dieser Gerüststrukturen Ring für Ring, die durch einen Regelsatz gesteuert wird, erzeugt einen Ast von Gerüststrukturen unterschiedlicher Größe. In dieser Hierarchie wird die kleinere Struktur als „Kind“, die größere als „Elter“ bezeichnet. Das Ergebnis der Prozessierung des Dictionary of Natural Products[5] ist ein Baumdiagramm, welches den chemischen Strukturraum der Naturstoffe darstellt. Jedes Molekül wird dabei, wie oben beschrieben, durch seine Gerüststruktur repräsentiert. Gerüststrukturen im selben Ast besitzen möglicherweise auch ähnliche Eigenschaften wie, zum Beispiel, ähnliche Bioaktivität.[89] Der Ausdruck des „Schwinghangelns“ bezeichnet dabei die Vereinfachung von Strukturen von außen nach innen entlang eines Astes bei ähnlicher biologischer Aktivität.

Basierend auf dem Fachwissen aus den Bereichen der Medizinalchemie und der organischen Synthese wurde in Zusammenarbeit mit Novartis ein neuer, generischer Regelsatz

entwickelt.[56] Diese Regelsatz erlaubt die Erzeugung von vielseitiger verwendbareren Baumdiagrammen chemischer Gerüststrukturen. So werden die Lücken im Baumdiagram durch „virtuelle Gerüststrukturen“, die nicht in Molekülen im Datensatz enthalten sind, geschlossen. Die Erzeugung der Baumdiagramm nach dem neuen Regelsatz ist zudem unabhängig von dem prozessierten Datensatz, so dass Strukturbäume verschiedener Verbindungsbibliotheken kombiniert werden können.

6.2.2 Scaffold Hunter

Obwohl die Strukturbäume den chemischen Strukturraum auf aussagekräftige und intuitive Weise darstellen, wurde ihre Verwendung zur Inspiration und Ausrichtung von organischen Synthesen durch die Verwendung statischer Bilder bisher stark eingeschränkt. Ein Bild kann lediglich eine beschränkte Menge an Information darstellen. Die Darstellung zusätzlicher, struktur-bezogener Informationen, z.B. biologischer Aktivität, ist auf diesem Wege kaum möglich. Daher wurde in Zusammenarbeit dem Lehrstuhl für Algorithmen an der Technischen Universität Dortmund ein Projekt entworfen und initiiert, in dessen Rahmen ein Computerprogramm namens “Scaffold Hunter” zur interaktiven Betrachtung chemischer Strukturbäume entwickelt werden sollte. Die Entwicklung und Betreuung des Projekts sowie die Anleitung der Projektgruppe von zwölf Informatikstudenten über ein Jahr erfolgte im Rahmen dieser Arbeit.

Scaffold Hunter visualisiert automatisch die Strukturbäume ausgehend von den Daten in einer Datenbank. Es ermöglicht die Navigation in den Strukturbäumen einschließlich der Anwendung von Filtern, zoomen, Farbverläufen basierend auf molekularen Eigenschaften und der Markierung bestimmter Strukturen. Eine Einführung in Scaffold Hunter findet sich im „Supporting Movie 2“ zu der entsprechenden Publikation in Nature Chemical Biology.[123] Die entsprechende Datenbank der Gerüststrukturen kann leicht mit Hilfe eines zweiten Programms, des „Scaffold Tree Generator“ erstellt werden. Dieses Programm wurde von S. Renner entwickelt und verarbeitet SD Dateien, ein offenes Standardformat für Molekülstrukturen. Der Scaffold Tree Generator wurde im Rahmen dieser Arbeit verändert, wobei im wesentlichen Fehler bereinigt und die Datenbankverbindung aktualisiert wurden.

6.2.3 Identifikation und Besetzung von Lücken im Chemischen Strukturraum

Wie oben beschrieben, ermöglicht der neue Regelsatz die Erzeugung von virtuellen Gerüststrukturen, die keine Verbindungen im Datensatz repräsentieren. Diese Strukturen vervollständigen den Strukturbaum und stellen somit Lücken im chemischen Strukturraum dar, der durch die analysierten Verbindungen abgedeckt wird. Solche Lücken können genutzt werden, um vielversprechende Gerüststrukturen auf Grund ihrer Nähe zu Gerüststrukturen von aktiven Verbindungen zu identifizieren. Ein erster Nachweis der Machbarkeit wurde erbracht durch eine Analyse bestehender Daten. Dazu wurden vielversprechende virtuelle

Gerüststrukturen aus den Daten in der PubChem Datenbank identifiziert. Nachfolgend wurden aktive Verbindungen, welche die entsprechenden Gerüststrukturen enthalten und ebenfalls als biologisch aktiv beschrieben sind, in der WOMBAT Datenbank gesucht. Die WOMBAT Datenbank enthält Moleküle und die dazugehörigen biologischen Aktivitäten, welche aus der wissenschaftlichen Literatur extrahiert wurden. Zur Identifikation neuer, aktiver Verbindungen wurden die Daten des Pyruvatkinase Datensatzes in PubChem analysiert. Von 65 vielversprechenden virtuellen Gerüststrukturen wurden vier ausgewählt und Verbindungsbibliotheken basierend auf diesen Strukturen gekauft. Das Testen dieser Verbindungen auf die Beeinflussung der Aktivität von Pyruvatkinase ergab acht potente Substanzen. Gemäß einer SciFinder Recherche war keine der entsprechenden Strukturen bis dato mit Pyruvatkinase in Verbindung gebracht worden. Damit konnte gezeigt werden, dass Scaffold Hunter Wissenschaftler in die Lage versetzt, ihre Daten in Wissen und letztlich in neue Wissenschaft umzusetzen. Scaffold Hunter selbst sowie die Validierungsstudien sind publiziert.[123] Alle Daten sind verfügbar; die Programme können kostenlos unter einer Lizenz für quelloffene Software von der Webseite www.scaffoldhunter.com heruntergeladen werden. Die genutzte Datenbank sowie eine Installationsanleitung sind vom edoc Server der Max-Planck Gesellschaft erhältlich: http://edoc.mpg.de/display.epl?mode=doc&id=429252.

6.2.4 Die Nutzung von Naturstoffen: γ-Pyrone

Eine offensichtliche Anwendung der Strukturbäume, die durch den neuen Regelsatz ermöglicht wird, ist der Vergleich verschiedener Verbindungsbibliotheken durch Kombination ihrer Strukturbäume. Dieser Ansatz kann auch genutzt werden, um eine der Bibliotheken mit Eigenschaften der anderen zu annotieren, beispielsweise mit Information über mögliche Zielproteine. Eine solche Information wäre von großem Wert bei der Entwicklung neuer Molekülbibliotheken. In einer ersten Anwendung wurde der chemische Strukturraum der Naturstoffe mit den Informationen über Zielproteine aus der WOMBAT Datenbank annotiert. Der Ast der γ-Pyrone zeigte vielversprechende biologische Aktivitäten und war synthetisch zugänglich. Darüber hinaus könnte das Testen von Verbindungen aus verschiedenen Hierarchieebenen dieses Astes Beispiele für Schwinghangeln im Segment der *O*-Heterocyclen aufzeigen. Für die Bereiche der Carbocyclen[23] sowie der *N*-Heterocyclen[44,48] wurden solche Beispiele bereits publiziert. Eine Substanzbibliothek höherer γ-Pyrone (mit Gerüststrukturen aus 2-4 Ringen) wurde zusammengestellt und gekauft, während eine Substanzbibliothek um die Wurzel das Astes, das γ-Pyron Gerüst durch Dipl.-Chem-. Wolfram Wilk und Dr. Samy Chammaa synthetisiert wurde. Die Analyse bereits vorhandener Daten aus der WOMBAT Datenbank ergab mehrere Fälle, in denen Schwinghangeln über drei Hierarchieebenen hinweg möglich war. Für die Identifikation neuer Inhibitoren wurden die Monoamine Oxidasen (MAO) A und B, die signal transducers and activators of transcription (STAT) Proteine sowie die saure Sphingomyelinase als Zielproteine ausgewählt, da für sie Aktivität für mindestens zwei

Gerüsttypen des γ-Pyron Astes beschrieben war. Die nachfolgende biochemische Evaluierung der Substanzbibliothek ergab mehrere, selektive Verbindungen für alle Zielproteine. Die MAO A und B Tests wurden zusammen mit Dipl.-Chem. Wolfram Wilk durchgeführt, während die Tests für die STATs und die Sphingomyelinase durch Kooperationspartner erfolgten. Insgesamt wurden 60 bzw. 35 aktive Substanzen für MAO A bzw. B gefunden, deren IC_{50} Werte sich zwischen einigen hundert nanomolar und 10 µM bewegten. Die Evaluierung gegen die STAT Proteine ergab zwei neue Inhibitoren, die zwar strukturell ähnlich aber vom Selektivitätsprofil her verschieden zu bekannten Inhibitoren waren. Eine der Verbindungen war auch in einem anspruchsvolleren Zell-basierten Test aktiv. Ebenfalls wurden drei Verbindungen mit guter Inhibition von saurer Sphingomyelinase and hoher Selektivität für das sauere Isoenzym gegenüber dem neutralen identifiziert. Diese Verbindungen besaßen außerdem wesentlich bessere molekulare Eigenschaften als die bekannten, nicht-peptidischen Inhibitoren. Diese Ergebnisse wurden publiziert[126] und belegen, dass Scaffold Hunter auch gut zur Annotation von Verbindungsbibliotheken mit möglichen Zielproteinen geeignet ist, wodurch zielgerichtetere Synthesen und biochemische Tests möglich sind. Gruppen von möglichen Zielproteinen für Gerüstfamilien können auch über PSSC gefunden werden, wie im folgenden Abschnitt beschrieben.

6.2.5 Ausblick

Dieser Abschnitt beschreibt zukünftige Erweiterungen von Scaffold Hunter, unter anderem verschiedene Methoden zur Generierung von Strukturbäumen an Hand individuell angepasster Regelsätze, einschließlich Biologie-basierter Strukturbäume, die von S. Renner entwickelt wurden.[89] Weitere Verbesserungen umfassen eine Substruktursuche, erweiterte Filtermöglichkeiten sowie eine Version, in der Regelsätze interaktiv verändert werden und die entstehenden Strukturbäume direkt visualisiert werden können. Die Anwendung von Scaffold Hunter zur Klassifikation andere Moleküle, z.B. von Katalysatoren, wird ebenso diskutiert wie die Erweiterung auf andere Anwendungsgebiete, z.B. zur Klassifikation von Proteinstrukturen. Schließlich wird eine mögliche Erweiterung zur Erzeugung chemisch sinnvoller Naturstofffragmente vorgeschlagen. Solche Fragmente könnten die Nutzung der chemischen Diversität in der Natur mit einem begrenzten synthetischen Aufwand ermöglichen.

6.3 Erforschung des Proteinstrukturraumes – Proteinstrukturähnlichkeitsclustering

Diese Projekt zielt darauf ab, die strukturellen Analogien zwischen Proteinen und den organischen Molekülen, welche die Proteinaktivität beeinflussen, zu nutzen, um Gruppen von potenziellen Zielproteinen für die biologische relevanten Moleküle zu identifizieren.

6.3.1 Stand und Ziele der Forschung

Die Bindungstaschen von Proteinen sowie die Moleküle, welche die Proteinaktivität beinflussen, müssen strukturell komplementär zu einander sein. Daher sollten strukturell ähnliche Bindungstaschen auf ähnliche Moleküle binden. Solches Verhalten ist literaturbekannt, z.B. von Kinasen. Die Nutzung der Ähnlichkeit auf atomarer Auflösung kann aus verschiedenen Gründen jedoch problematisch sein. So ist die Konformation eines Proteins nicht starr, da beispielsweise die Seitenketten der Aminosäuren flexibel sind. Auch kann die Auflösung von Kristallstrukturen experimentell begrenzt sein. Daher sollte die Ähnlichkeit auf einer abstrakteren Ebene definiert werden; auf der Ebene der Gerüststruktur. Als „Gerüststruktur“ wird in diesem Fall ein recht starres 3-dimensionales Gerüst bezeichnet, welches die Substituenten im Raum ausrichtet, die an den meisten molekularen Wechselwirkungen beteiligt sind. Solch ein Gerüst existiert sowohl in Proteinen wie auch in organischen Molekülen. In den Proteinen stellt das Peptid-Rückgrat, welches durch die Proteinfaltung räumlich angeordnet wird, das Gerüst dar, während es sich in organischen Molekülen meist um verbundene Ringsysteme handelt. Da die wechselwirkenden Substituenten in beiden Fällen trotz ähnlicher Gerüststrukturen sehr divers sein können, kann auf diesem Weg lediglich der Anfangspunkt zur Entwicklung eine Bibliothek diverser Moleküle gefunden werden, deren Diversität das Gegenstück zur Diversität der wechselwirkenden Proteinseitenketten bildet.

Anfangs basierte das Proteinstrukturähnlichkeitsclustering[43], das von Koch *et al*. genutzt wurde, auf dem manuellen Vergleich der Proteinfaltung mittels Datenbanken im Internet wie FSSP[192,195], CATH[244] oder SCOP.[132,153,154] Die Gruppen wurden von Hand erstellt durch Suche nach genereller Ähnlichkeit des Faltungstyps und anschließende visuelle Untersuchung der strukturellen Ähnlichkeit der Bindungstasche. Diese Vorgehensweise macht PSSC Analysen sehr aufwändig und subjektiv, da viele Parameter wie z.B. die Größe der Bindungstasche nicht präzise definiert waren.

Das Ziel dieser Arbeit war die Entwicklung eines definierten, automatisierten PSSC Prozesses und die Anwendung dieses Prozesses auf einen größeren Satz von Proteinstrukturen, idealerweise auf die gesamte Protein Data Bank (PDB). Die so gefunden Gruppen von Zielproteinen sollten dann experimentell validiert werden, um den Anwendungsbereich von PSSC zu erforschen.

6.3.2 Methodenentwicklung – automatisiertes PSSC

Der neu entwickelte PSSC Prozeß basiert auf der strukturellen Überlagerung von sogenannten Liganden-bindenden Kernstrukturen, d.h. kugelförmigen Ausschnitten aus der Struktur um die Bindungstasche herum, anstatt auf der Überlagerung von ganzen Proteinstrukturen. Diese Vorgehensweise führt zu einer drastischen Reduktion an falsch positiven Ergebnissen, in denen strukturelle Ähnlichkeit in Bereichen außerhalb der Bindungstasche vorhanden ist. Die Kriterien für die Erzeugung der Ligand-bindenden Kernstrukturen wurden ausgehend von einer Analyse

der katalytischen Zentren entwickelt, welche im Catalytic Site Atlas (CSA)[200] annotiert sind. Ein selbst entwickeltes Computerprogramm extrahierte die Ligand-bindenden Kernstrukturen aus den Strukturdaten in der PDB basierend auf der annotierten Bindungstasche aus dem CSA. Die nachfolgenden Strukturüberlagerungen wurden mit DaliLite[219,220] berechnet. Diese Ergebnisse sind mit denen früherer Analysen vergleichbar, da die verwendete FSSP Datenbank ebenfalls mit Dali erzeugt wurde. Die Ergebnisse der Strukturüberlagerungen wurden dann für ein statistisches Clustering verwendet, dessen Algorithmus in Java implementiert wurde. Alle Programme für den PSSC Prozeß wurden selbst geschrieben und auf hohe Leistung optimiert – mit Ausnahme von DaliLite.

Die Extraktion eines Satzes von Ligand-bindenden Kernstrukturen basierend auf dem CSA sowie eine Vorgruppierung nach struktureller Ähnlichkeit (RMSD < 1Å), welche die in der PDB vorhandene Redundanz reduziert, ergab 15,000 Ligand-bindende Kernstrukturen. Die Berechnung der Strukturüberlagerungen (von jede Kernstruktur gegen jede andere) benötigte zwei volle Monate auf einem Linux Cluster mit 128 Prozessoren. Die Ergebnisse wurden in einer SQL Datenbank zur weiteren Analyse gespeichert. Da der Dali Algorithmus Strukturüberlagerungen und strukturelle Ähnlichkeit nur für ähnliche Strukturen berechnen kann und daher keine vollständige Ähnlichkeitsmatrix generiert, können statistische Clustering-Algorithmen in diesem Fall nicht verwendet werden. Daher wurden die Daten mit einer angepassten Implementierung des OptiSim Clustering Algorithmus[225] prozessiert. Der Vergleich mit den Daten der SCOP Datenbank, des Goldstandards bezüglich struktureller Ähnlichkeit von Proteinen, ergab eine große Übereinstimmung der gefundenen Cluster mit der in SCOP definierten Ähnlichkeit. Für 33 Cluster sagte die SCOP Datenbank weniger als die Hälfte aller Cluster Mitglieder vorher.

Zur experimentellen Validierung wurden zwei Zielproteine eines Clusters, die Pyruvatkinase (PK) und die Dihydropteroat Synthetase (DHPS) ausgewählt. Eine Bibliothek von 740 Sulfanilamiden, einer bekannten Klasse von DHPS Inhibitoren, wurde auf PK Inhibition getestet. Bedauerlicherweise identifizierte der Test keine aktiven Moleküle, was einerseits an den gewählten Testbedingungen liegen könnte. Andererseits könnte dieses Ergebnis aber auch bedeuten, dass das PSSC Konzept in dem untersuchten Fall nicht greift. Eine abschließende Betrachtung der Anwendbarkeit von PSSC als Konzept ist daher auf Grund dieser Ergebnisse nicht möglich.

Zur PSSC Analyse großer Zahlen von Strukturen wurde eine schnelle Methode zum Strukturvergleich basiert auf „protein fingerprints“ entwickelt. Diese Methode nutzt eine abstrakte Repräsentation der Ligand-bindenden Kernstrukturen als Vektoren, welche einen Geschwindigkeitsgewinn von mehreren Größenordnungen gegenüber Dali bringt. Kreuzvalidierung zeigte eine mittelmäßige Übereinstimmung zwischen den Ergebnissen dieser Methode und denen von Dali für den Datensatz von 15.000 Ligand-bindenden Kernstrukturen.

Höchstwahrscheinlich liegt dies an der zu großen Einfachheit der gewählten „fingerprints“, welche die 3-dimensionale Anordnung der Sekundärstrukturelemente nur ungenügend abbilden. Die Implementierung eines Vier-Punkt-„fingerprints“ sollte dieses Problem beheben und die Genauigkeit verbessern. Darüber hinaus kann das benutze Ähnlichkeitsmaß, der sogenannte Tanimoto-Ähnlichkeitsindex, für alle Strukturvergleiche berechnet werden. Die so resultierende Ähnlichkeitsmatrix kann dann mit statistischen Clustering Methoden, z.B. in der freien Statistiksoftware R, analysiert werden.

6.3.3 Methodenentwicklung –PSSC mit dynamisierten Proteinstrukturen

Ein wesentlicher Nachteil von PSSC ist die Notwendigkeit einer Struktur des zu untersuchenden Proteins. Obwohl gezeigt wurde, dass PSSC auch mit Homologie-Modellen funktioniert[43], können Ligand-induzierte Konformationsänderungen, der sogenannte „induced fit“, ein Problem darstellen. Zusammen mit B.D. Charette aus Professor Berkowitz Gruppe in Lincoln, Nebraska wurde die Dynamisierung von PSSC durch die Simulation von Moleküldynamiken im Rahmen seines Forschungsaufenthalts in Dortmund untersucht. Es konnte erfolgreich gezeigt werden, dass Moleküldynamiksimulationen den Übergang von der apo Struktur zu einer Struktur mit gebundenem Liganden berechnen und so die induzierte Konformationsänderung simulieren können. Die resultierenden Strukturen führen zu einer Verbesserung der Ergebnisse entsprechender PSSC Analysen.[238]

6.3.4 Ausblick

Die Entwicklung komplexer Struktur“fingerprints“, ausgehend von den bereits diskutierten Ergebnissen, wird beschrieben. Eine derartige „fingerprint“-basierte Technologie würde den Vergleich eines Datensatzes von 67,000 Ligand-bindenden Kernstrukturen erlauben, die auf den Bindungstaschen der ReliBase Datenbank[284,285] beruhen. Die PSSC Analyse dieser Gruppe von Bindungstaschen, die das strukturell bekannte Proteom darstellen, könnte neue Einsichten in die strukturellen Verwandschaft zwischen Proteinen erbringen. Darüber hinaus könnte sie die Entwicklung neuer Inhibitoren für neue Zielproteine ermöglichen. Die experimentelle Validierung ein solchen großangelegten PSSC Analyse wäre ebenfalls notwendig, um den Anwendbarkeitsbereich von PSSC zu bestimmen. Eine Erweiterung der Methode um dynamisierte Strukturen mittels Moleküldynamiksimulationen ist sehr gut vorstellbar. Ebenso könnten automatisch erzeugte Strukturmodelle in Fällen genutzt werden, in denen keine experimentell bestimmten Proteinstrukturen zur Verfügung stehen.

6.4 Verschiedene Projekte

Während der hier beschriebenen Doktorarbeit wurden auch einige andere Projekte unterstützt. Diese Arbeiten schließen die NMR-unterstütze Konformationsanalyse mittels Kraftfeldrechnungen von Makrocyclen[297,298] ebenso ein wie Docking Simulationen von

Inhibitorstrukturen zur Erklärung der experimentell beobachteten Selektivitäten.[300-302,305] Weitere Arbeiten befaßten sich mit der Auswertung, statistischen Analyse und Qualitätskontrolle von biochemischen Tests sowie der nachfolgenden Optimierung des experimentellen Protokolls für diese Tests.

7 References

[1] Kihlberg, J., invited lecture at the Max-Planck Institute of Molecular Physiology.

[2] Kirkpatrick, P.; Ellis, C. Chemical space, *Nature (London, United Kingdom)* **2004**, *432*, 823.

[3] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Advanced Drug Delivery Reviews* **2001**, *46*, 3-26.

[4] Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective, *Medicinal Research Reviews* **1996**, *16*, 3-50.

[5] *Dictionary of Natural Products, v17.2 (2008), Chapmann & Hall / CRC Informa*, http://www.crcpress.com/

[6] Berdy, J. Bioactive microbial metabolites: A personal view, *Journal of Antibiotics* **2005**, *58*, 1-26.

[7] Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F., 3rd; Schenck, R. J.; Trippe, A. J. Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry, *J Org Chem* **2008**, *73*, 4443-51.

[8] Siegel, M. G.; Vieth, M. Drugs in other drugs: a new look at drugs as fragments, *Drug discovery today* **2007**, *12*, 71-9.

[9] Baber, J. C.; Feher, M. Predicting synthetic accessibility: application in drug discovery and development, *Mini-Reviews in Medicinal Chemistry* **2004**, *4*, 681-692.

[10] Feher, M.; Schmidt, J. M. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry, *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 218-227.

[11] Wetzel, S.; Schuffenhauer, A.; Roggo, S.; Ertl, P.; Waldmann, H. Cheminformatic analysis of natural products and their chemical space, *Chimia* **2007**, *61*, 355-360.

[12] Grabowski, K.; Schneider, G. Properties and architecture of drugs and natural products revisited, *Current Chemical Biology* **2007**, *1*, 115-127.

[13] Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration, *Nucleic Acids Research* **2006**, *34*, D668-D672.

[14] Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space, *Journal of Combinatorial Chemistry* **2001**, *3*, 157-166.

[15] Oprea, T. I.; Gottfries, J.; Sherbukhin, V.; Svensson, P.; Kuhler, T. C. Chemical information management in drug discovery: optimizing the computational and combinatorial chemistry interfaces, *Journal of Molecular Graphics & Modelling* **2000**, *18*, 512-524.

[16] Oprea, T. I.; Zamora, I.; Ungell, A.-L. Pharmacokinetically Based Mapping Device for Chemical Space Navigation, *Journal of Combinatorial Chemistry* **2002**, *4*, 258-266.

[17] Larsson, J.; Gottfries, J.; Bohlin, L.; Backlund, A. Expanding the ChemGPS Chemical Space with Natural Products, *Journal of Natural Products* **2005**, *68*, 985-991.

[18] Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. ChemGPS-NP: Tuned for Navigation in Biologically Relevant Chemical Space, *Journal of Natural Products* **2007**, *70*, 789-794.

[19] Rosen, J.; Gottfries, J.; Muresan, S.; Backlund, A.; Oprea, T. I. Novel Chemical Space Exploration via Natural Products, *Journal of Medicinal Chemistry* **2009**, *52*, 1953-1962.

[20] Rosen, J.; Loevgren, A.; Kogej, T.; Muresan, S.; Gottfries, J.; Backlund, A. ChemGPS-NPweb: chemical space navigation online, *Journal of Computer-Aided Molecular Design* **2009**, *23*, 253-259.

[21] Rosen, J.; Rickardson, L.; Backlund, A.; Gullbo, J.; Bohlin, L.; Larsson, R.; Gottfries, J. ChemGPS-NP mapping of chemical compounds for prediction of anticancer mode of action, *Qsar & Combinatorial Science* **2009**, *28*, 436-446.

[22] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks, *Journal of Medicinal Chemistry* **1996**, *39*, 2887-2893.

[23] Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP), *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 17272-17277.

[24] Burke, M. D.; Schreiber, S. L. A planning strategy for diversity-oriented synthesis, *Angewandte Chemie, International Edition* **2004**, *43*, 46-58.

[25] Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs, *Journal of medicinal chemistry* **1998**, *41*, 3325-9.

[26] Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery, *Science (Washington, D. C.)* **2000**, *287*, 1964-1969.

[27] Shah, A. V.; Walters, W. P.; Murcko, M. A. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules?, *Journal of Medicinal Chemistry* **1998**, *41*, 3314-3324.

[28] Walters, W. P.; Ajay; Murcko, M. A. Recognizing molecules with drug-like properties, *Current Opinion in Chemical Biology* **1999**, *3*, 384-387.

[29] Anon A decade of drug-likeness, *Nature Reviews Drug Discovery* **2007**, *6*, 853.

[30] Hann, M. M.; Leach, A. R.; Burrows, J. N.; Griffen, E. Lead discovery and the concepts of complexity and lead-likeness in the evolution of drug candidates, *Comprehensive Medicinal Chemistry II* **2006**, *4*, 435-458.

[31] Muresan, S.; Sadowski, J. Properties guiding drug- and lead-likeness, *Methods and Principles in Medicinal Chemistry* **2008**, *37*, 441-461.

[32] Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: "Target Fishing" using two- and three-dimensional molecular descriptors, *Journal of Medicinal Chemistry* **2006**, *49*, 6802-6810.

[33] Oprea, T. I. Pursuing leadlikeness in pharmaceutical research, *Joint Meeting on Medicinal Chemistry, Proceedings, Vienna, Austria, June 20-23, 2005* **2005**, 1-4.

[34] Rishton, G. M. Nonleadlikeness and leadlikeness in biochemical screening, *Drug Discovery Today* **2002**, *8*, 86-96.

[35] Rishton, G. M. Natural products as a robust source of new drugs and drug leads: past successes and present day issues, *American Journal of Cardiology* **2008**, *101*, 43D-49D.

[36] Zhang, M.-Q.; Wilkinson, B. Drug discovery beyond the 'rule-of-five', *Current Opinion in Biotechnology* **2007**, *18*, 478-488.

[37] Breinbauer, R.; Vetter, I. R.; Waldmann, H. From protein domains to drug candidates: Natural products as guiding principles in the design and synthesis of compound libraries, *Angewandte Chemie, International Edition* **2002**, *41*, 2878-2890.

[38] Mueller, G. Medicinal chemistry of target family-directed masterkeys, *Drug Discovery Today* **2003**, *8*, 681-691.

[39] Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries, *Journal of Chemical Information and Modeling* **2008**, *48*, 68-74.

[40] Madden Edward, A. The interaction between intellectual property and drug regulatory systems: global perspectives, *IDrugs : the investigational drugs journal* **2007**, *10*, 116-20.

[41] Southall, N. T.; Ajay Kinase Patent Space Visualization Using Chemical Replacements, *Journal of Medicinal Chemistry* **2006**, *49*, 2103-2109.

[42] Sheridan, R. P. The most common chemical replacements in drug-like compounds, *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 103-108.

[43] Koch, M. A.; Wittenberg, L.-O.; Basu, S.; Jeyaraj, D. A.; Gourzoulidou, E.; Reinecke, K.; Odermatt, A.; Waldmann, H. Compound library development guided by protein structure similarity clustering and natural product structure, *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 16721-16726.

[44] Noeren-Mueller, A.; Reis-Correa, I., Jr.; Prinz, H.; Rosenbaum, C.; Saxena, K.; Schwalbe, H. J.; Vestweber, D.; Cagna, G.; Schunk, S.; Schwarz, O.; Schiewe, H.; Waldmann, H. Discovery of protein phosphatase inhibitor classes by biology-oriented synthesis, *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 10606-10611.

[45] Barun, O.; Sommer, S.; Waldmann, H. Asymmetric solid-phase synthesis of 6,6-spiroketals, *Angewandte Chemie, International Edition* **2004**, *43*, 3195-3199.

[46] Kumar, K.; Waldmann, H. Synthesis of Natural Product Inspired Compound Collections, *Angewandte Chemie, International Edition* **2009**, *48*, 3224-3242.

[47] Meseguer, B.; Alonso-Diaz, D.; Griebenow, N.; Herget, T.; Waldmann, H. Natural product synthesis on polymeric supports-synthesis and biological evaluation of an indolactam library, *Angewandte Chemie, International Edition* **1999**, *38*, 2902-2906.

[48] Noeren-Mueller, A.; Wilk, W.; Saxena, K.; Schwalbe, H.; Kaiser, M.; Waldmann, H. Discovery of a new class of inhibitors of Mycobacterium tuberculosis protein tyrosine phosphatase B by biology-oriented synthesis, *Angewandte Chemie, International Edition* **2008**, *47*, 5973-5977, S5973/1-S5973/33.

[49] Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery, *Journal of Chemical Information and Modeling* **2007**, *47*, 342-353.

[50] Schuerer, S. C.; Tyagi, P.; Muskal, S. M. Prospective Exploration of Synthetically Feasible, Medicinally Relevant Chemical Space, *Journal of Chemical Information and Modeling* **2005**, *45*, 239-248.

[51] Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups, *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 374-380.

[52] Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual exploration of the small-molecule chemical universe below 160 D, *Angewandte Chemie, International Edition* **2005**, *44*, 1504-1508.

[53] Gorse, A.-D. Diversity in medicinal chemistry space, *Current Topics in Medicinal Chemistry (Sharjah, United Arab Emirates)* **2006**, *6*, 3-18.

[54] van Deursen, R.; Reymond, J.-L. Chemical space travel, *ChemMedChem* **2007**, *2*, 636-640.

[55] Henkel, T.; Brunne, R. M.; Muller, H.; Reichel, F. Statistical investigation into the structural complementarity of natural products and synthetic compounds, *Angewandte Chemie, International Edition* **1999**, *38*, 643-647.

[56] Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification, *Journal of Chemical Information and Modeling* **2007**, *47*, 47-58.

[57] Wessjohann, L. A.; Ruijter, E.; Garcia-Rivera, D.; Brandt, W. What can a chemist learn from nature's macrocycles? - A brief, conceptual view, *Molecular Diversity* **2005**, *9*, 171-186.

[58] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31-36.

[59] Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation, *Journal of Chemical Information and Computer Sciences* **1989**, *29*, 97-101.

[60] Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics, *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 493-500.

[61] Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent developments of the Chemistry Development Kit (CDK) - an open-source Java library for chemo- and bioinformatics, *Current Pharmaceutical Design* **2006**, *12*, 2111-2120.

[62] *CTFile formats*, Symyx Technologies **2007**, www.symyx.com/downloads/public/ctfile/ctfile.pdf

[63] *IsisDraw 2.5*, MDL Information Systems Inc. **2002**, www.mdl.com

[64] *ChemDraw 2006*, CambridgeSoft Corporation **2006**, www.cambridgesoft.com

[65] Veretnik, S.; Fink, J. L.; Bourne, P. E. Computational biology resources lack persistence and usability, *PLoS Computational Biology* **2008**, *4*, No pp given.

[66] Eades, P. Drawing Free Trees, *Bulletin of the Institute of Combinatorics and its Applications* **1992**, *5*, 10-36.

[67] Carriere, J.; Kazman, R. Research report. Interacting with huge hierarchies: beyond cone trees, *Information Visualization, 1995. Proceedings.* **1995**, 74-81.

[68] Lin, C.-C.; Yen, H.-C. On Balloon Drawings of Rooted Trees, *Journal of Graph Algorithms and Applications* **2007**, *11*, 431–452.

[69] Maggiora, G. M.; Johnson, M. A. Introduction to similarity in chemistry, **1990**, 1-13.

[70] Carr, R. A. E.; Congreve, M.; Murray, C. W.; Rees, D. C. Fragment-based lead discovery: leads by design, *Drug Discovery Today* **2005**, *10*, 987-992.

[71] Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery, *Journal of Medicinal Chemistry* **2008**, *51*, 3661-3680.

[72] Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned, *Nature Reviews Drug Discovery* **2007**, *6*, 211-219.

[73] Siegel Miles, G.; Vieth, M. Drugs in other drugs: a new look at drugs as fragments, *Drug discovery today* **2007**, *12*, 71-9.

[74] Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. *WOMBAT: world of molecular bioactivity*; Wiley-VCH, 2005; Vol. 23.

[75] The Universal Protein Resource (UniProt) 2009, *Nucleic Acids Research* **2009**, *37*, D169-D174.

[76] *Pyruvate kinase assay data in PubChem*, http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=361&loc=ea_ras

[77] Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries, *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 11473-11478.

[78] *ChemDiv Inc., 6605 Nancy Ridge Drive, San Diego, CA 92121 USA, chemdiv@chemdiv.com*, www.chemdiv.com

[79] *Aurora Fine Chemicals Ltd., Reininghausstr. 49, A-8020 Graz, Austria, E-mail: aurora@aurorafinechemicals.com*, http://www.aurorafinechemicals.com/

[80] *Enzymatic Assay of PYRUVATE KINASE (EC 2.7.1.40), Sigma Prod. No. P-1903*, Sigma-Aldrich http://www.sigmaaldrich.com/sigma/enzyme%20assay/p1903enz.pdf

[81] *SciFinder*, Chemical Abstract Service (CAS), Columbus, Ohio, USA **2007**, http://www.cas.org/products/scifindr/index.html

[82] Verheij, H. J. Leadlikeness and structural diversity of synthetic screening libraries, *Molecular Diversity* **2006**, *10*, 377-388.

[83] Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective, *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 1308-1315.

[84] Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries, *Angewandte Chemie, International Edition* **1999**, *38*, 3743-3748.

[85] Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. Ligand binding efficiency: trends, physical basis, and implications, *Journal of Medicinal Chemistry* **2008**, *51*, 2432-2438.

[86] Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Last 25 Years, *Journal of Natural Products* **2007**, *70*, 461-477.

[87] Schwarz, O.; Jakupovic, S.; Ambrosi, H.-D.; Haustedt, L. O.; Mang, C.; Mueller-Kuhrt, L. Natural Products in Parallel Chemistry - Novel 5-Lipoxygenase Inhibitors from BIOS-Based Libraries Starting from alpha -Santonin, *Journal of Combinatorial Chemistry* **2007**, *9*, 1104-1113.

[88] Tan, D. S. Diversity-oriented synthesis, *Chemical Biology* **2007**, *2*, 483-518.

[89] Renner, S.; van Otterlo, W. A. L.; Dominguez Seoane, M.; Moecklinghoff, S.; Hofmann, B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T. I.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-guided mapping and navigation of chemical space, *Nature Chemical Biology* **2009**, *advance online publication*, doi:10.1038/nchembio.188.

[90] Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mracec, M.; Oprea, T. I. WOMBAT and WOMBAT-PK: bioactivity databases for lead and drug discovery, *Chemical Biology* **2007**, *2*, 760-786.

[91] Oprea, T. I. Current trends in lead discovery: Are we looking for the appropriate properties?, *Journal of Computer-Aided Molecular Design* **2002**, *16*, 325-334.

[92] Renner, S.; van Otterlo, W.; Seoane, M. D.; Möcklinghoff, S.; Hofmann, B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-Guided Mapping and Navigation of Chemical Space, *Nature Chemical Biology* **2009**, accepted.

[93] Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An

Analysis of ALOGP and CLOGP Methods, *Journal of Physical Chemistry A* **1998**, *102*, 3762-3772.

[94] Bortolato, M.; Chen, K.; Shih, J. C. Monoamine oxidase inactivation: From pathophysiology to therapeutics, *Advanced Drug Delivery Reviews* **2008**, *60*, 1527-1533.

[95] Ward, R. J.; Lallemand, F.; de Witte, P.; Dexter, D. T. Neurochemical pathways involved in the protective effects of nicotine and ethanol in preventing the development of Parkinson's disease: Potential targets for the development of new therapeutic agents, *Progress in Neurobiology (Amsterdam, Netherlands)* **2008**, *85*, 135-147.

[96] Youdim, M. B. H.; Edmondson, D.; Tipton, K. F. The therapeutic potential of monoamine oxidase inhibitors, *Nature Reviews Neuroscience* **2006**, *7*, 295-309.

[97] Carcereri de Prati, A.; Ciampa, A. R.; Cavalieri, E.; Zaffini, R.; Darra, E.; Menegazzi, M.; Suzuki, H.; Mariotto, S. STAT1 as a new molecular target of anti-inflammatory treatment, *Current Medicinal Chemistry* **2005**, *12*, 1819-1828.

[98] Marchesini, N.; Hannun, Y. A. Acid and neutral sphingomyelinases: roles and mechanisms of regulation, *Biochemistry and Cell Biology* **2004**, *82*, 27-44.

[99] Schuchman, E. H.; Desnick, R. J. In *The Metabolic and Molecular Bases of Inherited Disease*; Scriver, C. R., Sly, W. S., Childs, B., Beaudet, A. L., Valle, D., Kinzler, K. W., Vogelstein, B., Eds.; McGraw Hill: New York, 2001, p 3589-3610.

[100] Bolasco, A.; Fioravanti, R.; Carradori, S. Recent development of monoamine oxidase inhibitors, *Expert Opinion on Therapeutic Patents* **2005**, *15*, 1763-1782.

[101] Novaroli, L.; Reist, M.; Favre, E.; Carotti, A.; Catto, M.; Carrupt, P.-A. Human recombinant monoamine oxidase B as reliable and efficient enzyme source for inhibitor screening, *Bioorganic & Medicinal Chemistry* **2005**, *13*, 6212-6217.

[102] Lim, C. P.; Cao, X. Structure, function, and regulation of STAT proteins, *Molecular BioSystems* **2006**, *2*, 536-550.

[103] Buettner, R.; Mora, L. B.; Jove, R. Activated STAT signaling in human tumors provides novel molecular targets for therapeutic intervention, *Clin Cancer Res* **2002**, *8*, 945-54.

[104] Wittig, I.; Groner, B. Signal transducer and activator of transcription 5 (STAT5), a crucial regulator of immune and cancer cells, *Curr Drug Targets Immune Endocr Metabol Disord* **2005**, *5*, 449-63.

[105] Berg, T. Signal transducers and activators of transcription as targets for small organic molecules, *ChemBioChem* **2008**, *9*, 2039-44.

[106] Fletcher, S.; Turkson, J.; Gunning, P. T. Molecular approaches towards the inhibition of the signal transducer and activator of transcription 3 (Stat3) protein, *ChemMedChem* **2008**, *3*, 1159-1168.

[107] Buettner, R.; Kortylewski, M.; Pardoll, D.; Yu, H.; Jove, R. STAT proteins as molecular targets for cancer therapy, *Signal Transducers and Activators of Transcription (STATs)* **2003**, 645-661.

[108] Costantino, L.; Barlocco, D. STAT3 as a target for cancer drug discovery, *Current Medicinal Chemistry* **2008**, *15*, 834-843.

[109] Chen, Z.; Han, Z. C. STAT3: a critical transcription activator in angiogenesis, *Medicinal Research Reviews* **2008**, *28*, 185-200.

[110] Mueller, J.; Sperl, B.; Reindl, W.; Kiessling, A.; Berg, T. Discovery of chromone-based inhibitors of the transcription factor STAT5, *ChemBioChem* **2008**, *9*, 723-727.

[111] Trajkovic, K.; Hsu, C.; Chiantia, S.; Rajendran, L.; Wenzel, D.; Wieland, F.; Schwille, P.; Brugger, B.; Simons, M. Ceramide triggers budding of exosome vesicles into multivesicular endosomes, *Science* **2008**, *319*, 1244-7.

[112] Marsh, M.; van Meer, G. Cell biology. No ESCRTs for exosomes, *Science* **2008**, *319*, 1191-2.

[113] Goggel, R.; Winoto-Morbach, S.; Vielhaber, G.; Imai, Y.; Lindner, K.; Brade, L.; Brade, H.; Ehlers, S.; Slutsky, A. S.; Schutze, S.; Gulbins, E.; Uhlig, S. PAF-mediated pulmonary edema: a new role for acid sphingomyelinase and ceramide, *Nat Med* **2004**, *10*, 155-60.

[114] Petrache, I.; Natarajan, V.; Zhen, L.; Medler, T. R.; Richter, A. T.; Cho, C.; Hubbard, W. C.; Berdyshev, E. V.; Tuder, R. M. Ceramide upregulation causes pulmonary cell apoptosis and emphysema-like disease in mice, *Nat Med* **2005**, *11*, 491-8.

[115] Teichgraber, V.; Ulrich, M.; Endlich, N.; Riethmuller, J.; Wilker, B.; De Oliveira-Munding, C. C.; van Heeckeren, A. M.; Barr, M. L.; von Kurthy, G.; Schmid, K. W.; Weller, M.; Tummler, B.; Lang, F.; Grassme, H.; Doring, G.; Gulbins, E. Ceramide accumulation mediates inflammation, cell death and infection susceptibility in cystic fibrosis, *Nat. Med.* **2008**, *14*, 382-91.

[116] Arenz, C.; Giannis, A. Synthesis of the First Selective Irreversible Inhibitor of Neutral Sphingomyelinase, *Angew Chem Int Ed Engl* **2000**, *39*, 1440-1442.

[117] Nara, F.; Tanaka, M.; Hosoya, T.; Suzuki-Konagai, K.; Ogita, T. Scyphostatin, a neutral sphingomyelinase inhibitor from a discomycete, Trichopeziza mollissima: taxonomy of the producing organism, fermentation, isolation, and physico-chemical properties, *J Antibiot (Tokyo)* **1999**, *52*, 525-30.

[118] Uchida, R.; Tomoda, H.; Dong, Y.; Omura, S. Alutenusin, a specific neutral sphingomyelinase inhibitor, produced by Penicillium sp. FO-7436, *J Antibiot (Tokyo)* **1999**, *52*, 572-4.

[119] Waschol owski, V.; Giannis, A. Sphingolactones: selective and irreversible inhibitors of neutral sphingomyelinase, *Angew Chem Int Ed Engl* **2006**, *45*, 827-30.

[120] Kolzer, M.; Arenz, C.; Ferlinz, K.; Werth, N.; Schulze, H.; Klingenstein, R.; Sandhoff, K. Phosphatidylinositol-3,5-Bisphosphate is a potent and selective inhibitor of acid sphingomyelinase, *Biol Chem* **2003**, *384*, 1293-8.

[121] Okudaira, C.; Ikeda, Y.; Kondo, S.; Furuya, S.; Hirabayashi, Y.; Koyano, T.; Saito, Y.; Umezawa, K. Inhibition of acidic sphingomyelinase by xanthone compounds isolated from Garcinia speciosa, *Journal of Enzyme Inhibition* **2000**, *15*, 129-138.

[122] Hamada, M.; Iikubo, K.; Ishikawa, Y.; Ikeda, A.; Umezawa, K.; Nishiyama, S. Biological activities of alpha -mangostin derivatives against acidic sphingomyelinase, *Bioorganic & Medicinal Chemistry Letters* **2003**, *13*, 3151-3153.

[123] Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive exploration of chemical space with Scaffold Hunter, *Nature Chemical Biology* **2009**, *advance online publication*, doi:10.1038/nchembio.187.

[124] Lee, M. L.; Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: Application in the design of natural product-based combinatorial libraries, *Journal of Combinatorial Chemistry* **2001**, *3*, 284-289.

[125] Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive Exploration of Chemical Space with Scaffold Hunter, *Nature Chemical Biology* **2009**, in press.

[126] Wetzel, S.; Wilk, W.; Chamma, S.; Sperl, B.; Roth, A.; Renner, S.; Berg, T.; Arenz, C.; Giannis, A.; Oprea, T. I.; Rauh, D.; Kaiser, M.; Waldmann, H. Prospective Bioactivity Annotation by Scaffold Tree Merging, *Proceedings of the National Academy of Sciences of the United States of America* **2009**, in revision.

[127] Correa, I. R.; Noren-Muller, A.; Ambrosi, H. D.; Jakupovic, S.; Saxena, K.; Schwalbe, H.; Kaiser, M.; Waldmann, H. Identification of inhibitors for mycobacterial protein tyrosine

phosphatase B (MptpB) by biology-oriented synthesis (BIOS), *Chemistry-an Asian Journal* **2007**, *2*, 1109-1126.

[128] Arve, L.; Voigt, T.; Waldmann, H. Charting biological and chemical space: PSSC and SCONP as guiding principles for the development of compound collections based on natural product scaffolds, *Qsar & Combinatorial Science* **2006**, *25*, 449-456.

[129] Dekker, F. J.; Wetzel, S.; Waldmann, H. Natural product scaffolds and protein structure similarity clustering (PSSC) as inspiration sources for compound library design in chemogenomics and drug development, *Chemogenomics* **2006**, 59-84.

[130] *PubChem - a component of NIH's Molecular Libraries Roadmap Initiative*, **2009**, http://pubchem.ncbi.nlm.nih.gov/

[131] Steinbeck, C.; Al-Lazikani, B.; Hermjakob, H.; Overington, J.; Thornton, J. New open drug activity data at EBI, *Chemistry Central Journal* **2009**, *3*, O3.

[132] Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology* **1995**, *247*, 536-40.

[133] Erlanson, D. A. Fragment-based lead discovery: a chemical update, *Current Opinion in Biotechnology* **2006**, *17*, 643-652.

[134] Murray, C. W.; Rees, D. C. The rise of fragment-based drug discovery, *Nat Chem* **2009**, *1*, 187-192.

[135] Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-based lead discovery, *Nat Rev Drug Discov* **2004**, *3*, 660-672.

[136] Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine, *Nature (London, United Kingdom)* **2004**, *432*, 855-861.

[137] Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery, *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 856-864.

[138] Keseru, G. M.; Makara, G. M. The influence of lead discovery strategies on the properties of drug candidates, *Nat Rev Drug Discov* **2009**, *8*, 203-212.

[139] Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for fragment-based lead discovery?, *Drug Discovery Today* **2003**, *8*, 876-877.

[140] Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry, *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 511-522.

[141] Hubbard, R. E.; Chen, I.; Davis, B. Informatics and modeling challenges in fragment-based drug discovery, *Current Opinion in Drug Discovery and Development* **2007**, *10*, 289-297.

[142] *Enzymatic Assay of L-LACTIC DEHYDROGENASE1 (EC 1.1.1.27)*, Sigma-Aldrich http://www.sigmaaldrich.com/sigma/enzyme%20assay/l1254enz.pdf

[143] Schust, J.; Sperl, B.; Hollis, A.; Mayer, T. U.; Berg, T. Stattic: A Small-Molecule Inhibitor of STAT3 Activation and Dimerization, *Chemistry & Biology (Cambridge, MA, United States)* **2006**, *13*, 1235-1242.

[144] Schust, J.; Berg, T. A high-throughput fluorescence polarization assay for signal transducer and activator of transcription 3, *Analytical Biochemistry* **2004**, *330*, 114-118.

[145] Mueller, J.; Schust, J.; Berg, T. A high-throughput assay for signal transducer and activator of transcription 5b based on fluorescence polarization, *Analytical Biochemistry* **2008**, *375*, 249-254.

[146] *Proteins - Structure and Function*; Whitford, D., Ed.; John Wiley & Sons: Chichester, UK, 2005.

[147] Yeats, C. A.; Orengo, C. A. In *Handbook of Proteins*; Cox, M. M., Phillips, G. N. J., Eds.; John Wiley & Sons: Chicherster, UK, 2007; Vol. 1, p 23-32.

[148] Coulson Andrew, F. W.; Moult, J. A unifold, mesofold, and superfold model of protein fold use, *Proteins* **2002**, *46*, 61-71.

[149] Grant, A.; Lee, D.; Orengo, C. Progress towards mapping the universe of protein folds, *Genome biology* **2004**, *5*, 107.

[150] Koonin, E. V.; Wolf, Y. I.; Karev, G. P. The structure of the protein universe and genome evolution, *Nature (London, United Kingdom)* **2002**, *420*, 218-223.

[151] Leonov, H.; Mitchell, J. S. B.; Arkin, I. T. Monte Carlo estimation of the number of possible protein folds: Effects of sampling bias and folds distributions, *Proteins: Structure, Function, and Genetics* **2003**, *51*, 352-359.

[152] Sadreyev Ruslan, I.; Grishin Nick, V. Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds, *BMC structural biology* **2006**, *6*, 6.

[153] Andreeva, A.; Howorth, D.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. SCOP database in 2004: refinements integrate structure and sequence family data, *Nucleic Acids Research* **2004**, *32*, D226-D229.

[154] Andreeva, A.; Howorth, D.; Chandonia, J.-M.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. Data growth and its impact on the SCOP database: new developments, *Nucleic Acids Research* **2008**, *36*, D419-D425.

[155] *statistics taken from the SCOP database website*, http://scop.mrc-lmb.cam.ac.uk/scop/count.htmlscop-1.75

[156] Andreeva, A.; Murzin, A. G. Evolution of protein fold in the presence of functional constraints, *Current Opinion in Structural Biology* **2006**, *16*, 399-408.

[157] Grishin, N. V. Fold change in evolution of protein structures, *Journal of structural biology* **2001**, *134*, 167-85.

[158] Taylor, W. R. Evolutionary transitions in protein fold space, *Current Opinion in Structural Biology* **2007**, *17*, 354-361.

[159] Balamurugan, R.; Dekker, F. J.; Waldmann, H. Design of compound libraries based on natural product scaffolds and protein structure similarity clustering (PSSC), **2005**, *1*, 36-45.

[160] Dekker, F. J.; Wetzel, S.; Waldmann, H. Natural product scaffolds and protein structure similarity clustering (PSSC) as inspiration sources for compound library design in chemogenomics and drug development, **2006**, 59-84.

[161] Koch, M. A.; Waldmann, H. Natural product-derived compounds libraries and protein structure similarity as guiding principles for the discovery of drug candidates, **2004**, *22*, 377-403.

[162] Koch, M. A.; Waldmann, H. Protein domain fold similarity and natural product structure as guiding principles for compound library design, **2005**, *51*, 1-18.

[163] Koch, M. A.; Wittenberg, L.-O.; Basu, S.; Jeyaraj, D. A.; Gourzoulidou, E.; Reinecke, K.; Odermatt, A.; Waldmann, H. Compound library development guided by protein structure similarity clustering and natural product structure, **2004**, *101*, 16721-16726.

[164] Jones, S.; Thornton, J. M. Searching for functional sites in protein structures, *Current Opinion in Chemical Biology* **2004**, *8*, 3-7.

[165] Pettit, F. K.; Bare, E.; Tsai, A.; Bowie, J. U. HotPatch: A Statistical Approach to Finding Biologically Relevant Features on Protein Surfaces, *Journal of Molecular Biology* **2007**, *369*, 863-879.

[166] Russell, R. B.; Sasieni, P. D.; Sternberg, M. J. E. Supersites within superfolds. Binding site similarity in the absence of homology, *Journal of Molecular Biology* **1998**, *282*, 903-918.

[167] Stark, A.; Shkumatov, A.; Russell, R. B. Finding Functional Sites in Structural Genomics Proteins, *Structure (Cambridge, MA, United States)* **2004**, *12*, 1405-1412.

[168] Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead discovery using molecular docking, *Current Opinion in Chemical Biology* **2002**, *6*, 439-446.

[169] Lang, P. T.; Aynechi, T.; Moustakas, D.; Shoichet, B.; Kuntz, I. D.; Brooijmans, N.; Oshiro, C. M. Molecular docking and structure-based design, *Drug Discovery Research* **2007**, 3-23, 1 plate.

[170] Andrusier, N.; Mashiach, E.; Nussinov, R.; Wolfson Haim, J. Principles of flexible protein-protein docking, *Proteins* **2008**, *73*, 271-89.

[171] Dias, R.; Filgueira de Azevedo, W., Jr. Molecular docking algorithms, *Current Drug Targets* **2008**, *9*, 1040-1047.

[172] Morris, G. M.; Lim-Wilby, M. Molecular docking, *Methods in Molecular Biology (Totowa, NJ, United States)* **2008**, *443*, 365-382.

[173] Breda, A.; Basso, L. A.; Santos, D. S.; de Azevedo, W. F., Jr. Virtual screening of drugs: score functions, docking, and drug design, *Current Computer-Aided Drug Design* **2008**, *4*, 265-272.

[174] Rajamani, R.; Good, A. C. Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development, *Current Opinion in Drug Discovery & Development* **2007**, *10*, 308-315.

[175] Schulz-Gasch, T.; Stahl, M. Scoring functions for protein-ligand interactions: a critical perspective, *Drug Discovery Today: Technologies* **2004**, *1*, 231-239.

[176] Spyrakis, F.; Kellogg, G. E.; Amadasi, A.; Cozzini, P. Scoring functions for virtual screening, *Frontiers in Drug Design and Discovery* **2007**, *3*, 317-379.

[177] Stouten, P. F. W.; Kroemer, R. T. Docking and scoring, *Comprehensive Medicinal Chemistry II* **2006**, *4*, 255-281.

[178] Warren, G. L.; Peishoff, C. E.; Head, M. S. Docking algorithms and scoring functions; state-of-the-art and current limitations, *Computational and Structural Approaches to Drug Discovery* **2008**, 137-154.

[179] Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions, *Journal of Medicinal Chemistry* **2006**, *49*, 5912-5931.

[180] Johnson, A. P.; Valko, V.; Valko, A.; Zsoldos, Z.; Boda, K.; Reid, D. SynSPROUT and SPROUT-LeadOpt: De novo ligand design and optimization guided by virtual synthesis, *Abstracts of Papers, 233rd ACS National Meeting, Chicago, IL, United States, March 25-29, 2007* **2007**, COMP-037.

[181] Krueger, B. A.; Dietrich, A.; Baringhaus, K.-H.; Schneider, G. Scaffold-hopping potential of fragment-based de novo design: the chances and limits of variation, *Combinatorial Chemistry & High Throughput Screening* **2009**, *12*, 383-396.

[182] Law, J. M. S.; Fung, D. Y. K.; Zsoldos, Z.; Simon, A.; Szabo, Z.; Csizmadia, I. G.; Johnson, A. P. Validation of the SPROUT de novo design program, *Theochem* **2003**, *666-667*, 651-657.

[183] Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-Based Approach to de Novo Design Using Reaction Vectors, *Journal of Chemical Information and Modeling* **2009**, *49*, 1163-1184.

[184] Gerlach, C.; Muenzel, M.; Baum, B.; Gerber, H.-D.; Craan, T.; Diedrich, W. E.; Klebe, G. KNOBLE: a knowledge-based approach for the design and synthesis of readily accessible small-molecule chemical probes to test protein binding, *Angewandte Chemie, International Edition* **2007**, *46*, 9105-9109.

[185] Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and application, *Journal of Chemical Information and Modeling* **2007**, *47*, 279-294.

[186] Schmitt, S.; Hendlich, M.; Klebe, G. From structure to function: A new approach to detect functional similarity among proteins independent from sequence and fold homology, *Angewandte Chemie, International Edition* **2001**, *40*, 3141-3144.

[187] Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology, *Journal of Molecular Biology* **2002**, *323*, 387-406.

[188] Kuhn, D.; Weskamp, N.; Schmitt, S.; Huellermeier, E.; Klebe, G. From the Similarity Analysis of Protein Cavities to the Functional Classification of Protein Families Using Cavbase, *Journal of Molecular Biology* **2006**, *359*, 1023-1044.

[189] Kuhn, D.; Weskamp, N.; Huellermeier, E.; Klebe, G. Functional classification of protein kinase binding sites using Cavbase, *ChemMedChem* **2007**, *2*, 1432-1447.

[190] Koch, M. A.; Waldmann, H. Protein structure similarity clustering and natural product structure as guiding principles in drug discovery, *Drug Discovery Today* **2005**, *10*, 471-483.

[191] Lamb, S. S.; Wright, G. D. Accessorizing natural products: adding to nature's toolbox, *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 519-520.

[192] Holm, L.; Sander, C. The FSSP database of structurally aligned protein fold families, *Nucleic Acids Research* **1994**, *22*, 3600-9.

[193] Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Engineering* **1998**, *11*, 739-747.

[194] Shindyalov, I. N.; Bourne, P. E. A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm, *Nucleic Acids Research* **2001**, *29*, 228-229.

[195] Holm, L.; Sander, C. Dali/FSSP classification of three-dimensional protein folds, *Nucleic Acids Research* **1997**, *25*, 231-234.

[196] Holm, L.; Sander, C. Protein structure comparison by alignment of distance matrixes, *Journal of Molecular Biology* **1993**, *233*, 123-38.

[197] Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C. Selection of representative protein data sets, *Protein Science* **1992**, *1*, 409-17.

[198] Hobohm, U.; Sander, C. Enlarged representative set of protein structures, *Protein Science* **1994**, *3*, 522-4.

[199] Shindyalov, I. N.; Bourne, P. E. Protein sequence-structure space and resultant data redundancy in the protein data bank, *METMBS '01, Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, Las Vegas, NV, United States, June 25-28, 2001* **2001**, 139-145.

[200] Porter, C. T.; Bartlett, G. J.; Thornton, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Research* **2004**, *32*, D129-D133.

[201] *statistics taken from the Catalytic Site Atlas website*, http://www.ebi.ac.uk/thornton-srv/databases/CSA/

[202] Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **1983**, *22*, 2577-637.

[203] Bourne, P. E.; Shindyalov, I. N. Structure comparison and alignment, *Methods of Biochemical Analysis* **2003**, *44*, 321-337.

[204] Kolodny, R.; Petrey, D.; Honig, B. Protein structure comparison: Implications for the nature of 'fold space', and structure and function prediction, *Current Opinion in Structural Biology* **2006**, *16*, 393-398.

[205] Sillitoe, I.; Orengo, C. Protein structure comparison, *Bioinformatics* **2003**, 81-101.

[206] Madej, T.; Gibrat, J.-F.; Bryant, S. H. Threading a database of protein cores, *Proteins: Structure, Function, and Genetics* **1995**, *23*, 356-69.

[207] Bhattacharya, S.; Bhattacharyya, C.; Chandra, N. R. Comparison of protein structures by growing neighborhood alignments, *BMC Bioinformatics* **2007**, *8*, No pp given.

[208] Bostick, D.; Vaisman, I. I. A new topological method to measure protein structure similarity, *Biochemical and Biophysical Research Communications* **2003**, *304*, 320-325.

[209] Chu, C.-H.; Tang, C. Y.; Tang, C.-Y.; Pai, T.-W. Angle-distance image matching techniques for protein structure comparison, *Journal of Molecular Recognition* **2008**, *21*, 442-452.

[210] Eslahchi, C.; Pezeshk, H.; Sadeghi, M.; Massoud Rahimi, A.; Maboudi Afkham, H.; Arab, S. STON: A novel method for protein three-dimensional structure comparison, *Computers in Biology and Medicine* **2009**, *39*, 166-172.

[211] Fang, H.; Xiang, J.; Hu, M. A fast approach to protein structure alignment, *Journal of Computational and Theoretical Nanoscience* **2007**, *4*, 1369-1374.

[212] Krissinel, E.; Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallographica, Section D: Biological Crystallography* **2004**, *D60*, 2256-2268.

[213] Pelta, D. A.; Gonzalez, J. R.; Vega, M. M. A simple and fast heuristic for protein structure comparison, *BMC Bioinformatics* **2008**, *9*, No pp given.

[214] Plewczynski, D.; Pas, J.; Von Grotthuss, M.; Rychlewski, L. 3D-Hit: fast structural comparison of proteins, *Applied Bioinformatics* **2002**, *1*, 223-225.

[215] Thompson, K. E. *Improving VAST structure alignment performance and analysis of small molecule contacts in protein structures*; doctoral thesis; Johns Hopkins Univ., Baltimore, MD, USA.; 2008.

[216] Tsang, H. S. *Vector alignment search tool (VAST) automated protein structure comparison using special structural elements*; doctoral thesis; Johns Hopkins Univ., Baltimore, MD, USA; 2007.

[217] Wu, Z.; Wang, Y.; Feng, E.; Chen, L. A new geometric-topological method to measure protein fold similarity, *Chemical Physics Letters* **2007**, *433*, 432-438.

[218] Zotenko, E.; O'Leary, D. P.; Przytycka, T. M. Secondary structure spatial conformation footprint: a novel method for fast protein structure comparison and classification, *BMC structural biology* **2006**, *6*, No pp given.

[219] Holm, L.; Park, J. DaliLite workbench for protein structure comparison, *Bioinformatics* **2000**, *16*, 566-567.

[220] Holm, L.; Kaariainen, S.; Rosenstrom, P.; Schenkel, A. Searching protein structure database with DaliLite v.3, *Bioinformatics* **2008**, *24*, 2780-2781.

[221] *VAST search webpage*, National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html

[222] Sierk, M. L.; Pearson, W. R. Sensitivity and selectivity in protein structure comparison, *Protein Science* **2004**, *13*, 773-785.

[223] *DALILite Download webpage*, http://ekhidna.biocenter.helsinki.fi/dali_lite/downloads/v3/

[224] Jain, A. K.; Murty, M. N.; Flynn, P. J. Data Clustering: A Review, *ACM Computing Surveys* **1999**, *31*, 264-323.

[225] Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets, *Journal of Chemical Information and Computer Sciences* **1997**, *37*, 1181-1188.

[226] *Pipeline Pilot, a workflow tool*, Accelrys http://accelrys.com/products/scitegic/

[227] Willett, P. Similarity-based approaches to virtual screening, *Biochemical Society Transactions* **2003**, *31*, 603-606.

[228] Willett, P. Similarity-based virtual screening using 2D fingerprints, *Drug discovery today* **2006**, *11*, 1046-1053.

[229] Gunasekera, S. P.; McCarthy, P. J.; Kelly-Borges, M.; Lobkovsky, E.; Clardy, J. Dysidiolide: a novel protein phosphatase inhibitor from the Caribbean sponge Dysidea etheria de Laubenfels, *Journal of the American Chemical Society* **1996**, *118*, 8759-8760.

[230] Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing, *Proteins: Structure, Function, and Genetics* **1990**, *8*, 195-202.

[231] Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; VanGunsteren, W. F.; Mark, A. E. Peptide folding: when simulation meets experiment, *Angewandte Chemie, International Edition* **1999**, *38*, 236-240.

[232] DaCosta, S. A.; Schumaker, L. M.; Ellis, M. J. Mannose 6-phosphate/insulin-like growth factor 2 receptor, a bona fide tumor suppressor gene or just a promising candidate?, *Journal of mammary gland biology and neoplasia* **2000**, *5*, 85-94.

[233] Jirtle, R. L. Multifaceted M6P/IGF2R liver tumor suppressor, *Normal and Malignant Liver Cell Growth, Proceedings of the International Falk Workshop, Halle, Germany, Jan. 29-30, 1998* **1999**, 136-140.

[234] Oates, A. J.; Schumaker, L. M.; Jenkins, S. B.; Pearce, A. A.; Dacosta, S. A.; Arun, B.; Ellis, M. J. C. The mannose 6-phosphate/insulin-like growth factor 2 receptor (M6P/IGF2R), a putative breast tumor suppressor gene, *Breast Cancer Research and Treatment* **1998**, *47*, 269-281.

[235] Siegenthaler, G.; Hotz, R.; Chatellard-Gruaz, D.; Jaconi, S.; Saurat, J. H. Characterization and expression of a novel human fatty acid-binding protein: The epidermal type (E-FABP), *Biochemical and Biophysical Research Communications* **1993**, *190*, 482-7.

[236] Gutierrez-Gonzalez, L. H.; Ludwig, C.; Hohoff, C.; Rademacher, M.; Hanhoff, T.; Ruterjans, H.; Spener, F.; Lucke, C. Solution structure and backbone dynamics of human epidermal-type fatty acid-binding protein (E-FABP), *Biochemical Journal* **2002**, *364*, 725-737.

[237] Sinha, P.; Hutter, G.; Kottgen, E.; Dietel, M.; Schadendorf, D.; Lage, H. Increased expression of epidermal fatty acid binding protein, cofilin, and 14-3-3-s (stratifin) detected by two-dimensional gel electrophoresis, mass spectrometry and microsequencing of drug-resistant human adenocarcinoma of the pancreas, *Electrophoresis* **1999**, *20*, 2952-2960.

[238] Charette, B. D.; MacDonald, R. G.; Wetzel, S.; Berkowitz, D. B.; Waldmann, H. Protein structure similarity clustering: dynamic treatment of PDB structures facilitates clustering, *Angewandte Chemie, International Edition* **2006**, *45*, 7766-7770.

[239] Bartlett, G. J.; Porter, C. T.; Borkakoti, N.; Thornton, J. M. Analysis of Catalytic Residues in Enzyme Active Sites, *Journal of Molecular Biology* **2002**, *324*, 105-121.

[240] Torrance, J. W.; Bartlett, G. J.; Porter, C. T.; Thornton, J. M. Using a Library of Structural Templates to Recognise Catalytic Sites and Explore their Evolution in Homologous Families, *Journal of Molecular Biology* **2005**, *347*, 565-581.

[241] Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research* **1997**, *25*, 3389-3402.

[242] Henikoff, S.; Henikoff, J. G. Embedding strategies for effective use of information from multiple sequence alignments, *Protein Science* **1997**, *6*, 698-705.

[243] Ferre, F. From sequence to structure: an easy approach to protein structure prediction, *Internet for Cell and Molecular Biologists (2nd Edition)* **2004**, 233-307.

[244] Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH - a hierarchic classification of protein domain structures, *Structure (London)* **1997**, *5*, 1093-1108.

[245] Wang, J.; Feng, J.-A. NdPASA: a novel pairwise protein sequence alignment algorithm that incorporates neighbor-dependent amino acid propensities, *Proteins: Structure, Function, and Bioinformatics* **2005**, *58*, 628-637.

[246] Zhu, J.; Weng, Z. FAST: a novel protein structure alignment algorithm, *Proteins: Structure, Function, and Bioinformatics* **2005**, *58*, 618-627.

[247] Paccanaro, A.; Casbon, J. A.; Saqi, M. A. S. Spectral clustering of protein sequences, *Nucleic Acids Research* **2006**, *34*, 1571-1580.

[248] Munoz, M. E.; Ponce, E. Pyruvate kinase: current status of regulatory and functional properties, *Comparative Biochemistry and Physiology, Part B: Biochemistry & Molecular Biology* **2003**, *135B*, 197-218.

[249] Uyeda, K. Pyruvate kinase, *Encyclopedia of Biological Chemistry* **2004**, *3*, 616-619.

[250] Mazurek, S. Pyruvate kinase type M2: a key regulator within the tumour metabolome and a tool for metabolic profiling of tumours, *Ernst Schering Foundation Symposium Proceedings* **2008**, 99-124.

[251] Zhang, X.; Ye, H. Research progress on relationship of M2-pyruvate kinase and cervical cancer, *Xiandai Shengwuyixue Jinzhan* **2008**, *8*, 1356-1357, 1324.

[252] Ferone, R. Folate metabolism in malaria, *Bulletin of the World Health Organization* **1977**, *55*, 291-8.

[253] Vinnicombe, H. G.; Derrick, J. P. Dihydropteroate synthase: an old drug target revisited, *Biochemical Society Transactions* **1999**, *27*, 53-58.

[254] Cravo, P.; Culleton, R.; Afonso, A.; Ferreira, I. D.; do Rosario, V. E. Mechanisms of drug resistance in malaria: current and new challenges, *Anti-Infective Agents in Medicinal Chemistry* **2006**, *5*, 63-73.

[255] Swarbrick, J.; Iliades, P.; Simpson, J. S.; Macreadie, I. Folate biosynthesis - reappraisal of old and novel targets in the search for new antimicrobials, *Open Enzyme Inhibition Journal* **2008**, *1*, 12-33.

[256] Alifrangis, M.; Enosse, S.; Khalil, I. F.; Tarimo, D. S.; Lemnge, M. M.; Thompson, R.; Bygbjerg, I. C.; Ronn, A. M. Prediction of Plasmodium falciparum resistance to sulfadoxine/pyrimethamine in vivo by mutations in the dihydrofolate reductase and

dihydropteroate synthetase genes: A comparative study between sites of differing endemicity, *American Journal of Tropical Medicine and Hygiene* **2003**, *69*, 601-606.

[257] A-Elbasit, I. E.; Alifrangis, M.; Khalil, I. F.; Bygbjerg, I. C.; Masuadi, E. M.; Elbashir, M. I.; Giha, H. A. The implication of dihydrofolate reductase and dihydropteroate synthetase gene mutations in modification of Plasmodium falciparum characteristics, *Malaria Journal* **2007**, *6*, No pp given.

[258] Christov, L. Xylanases: properties and applications, *Concise Encyclopedia of Bioresource Technology* **2004**, 601-609.

[259] Collins, T.; Gerday, C.; Feller, G. Xylanases, xylanase families and extremophilic xylanases, *FEMS Microbiology Reviews* **2005**, *29*, 3-23.

[260] Ming, H.; Nie, G. Purification and characterization of xylanase, *Xinxiang Yixueyuan Xuebao* **2007**, *24*, 308-310.

[261] Paice, M.; Renaud, S.; Bourbonnais, R.; Labonte, S.; Berry, R. Specificity of kraft pulp bleaching with xylanase - a key factor in cost benefit analysis, *TAPPI Engineering, Pulping and Environmental Conference, Philadelphia, PA, United States, Aug. 28-31, 2005* **2005**, Paice/1-Paice/12.

[262] Selvaraj, V.; Rajendran, A.; Thangavelu, V. Role of xylanase enzyme in paper & pulp industries, *Chemical Engineering World* **2007**, *42*, 108-110.

[263] Bertino, J. R.; Hillcoat, B. L. Regulation of dihydrofolate reductase and other folate-requiring enzymes, *Advances in Enzyme Regulation* **1968**, *6*, 335-49.

[264] Matthews, R. G. Mammalian methylenetetrahydrofolate reductase, **1991**, *1*, 371-87.

[265] Matthews, R. G.; Sheppard, C.; Goulding, C. Methylenetetrahydrofolate reductase and methionine synthase: biochemistry and molecular biology, *European Journal of Pediatrics* **1998**, *157*, S54-S59.

[266] Fodinger, M.; Horl, W. H.; Sunder-Plassmann, G. Molecular biology of 5,10-methylenetetrahydrofolate reductase, *Journal of nephrology* **2000**, *13*, 20-33.

[267] Jongbloet, P. H.; Verbeek, A. L. M.; den Heijer, M.; Roeleveld, N. Methylenetetrahydrofolate reductase (MTHFR) gene polymorphisms resulting in suboptimal oocyte maturation: a discussion of folate status, neural tube defects, schizophrenia, and vasculopathy, *Journal of Experimental & Clinical Assisted Reproduction* **2008**, *5*, No pp given.

[268] Stankova, J.; Lawrance, A. K.; Rozen, R. Methylenetetrahydrofolate reductase (MTHFR): a novel target for cancer therapy, *Current Pharmaceutical Design* **2008**, *14*, 1143-1150.

[269] Thomas, P.; Fenech, M. Methylenetetrahydrofolate reductase, common polymorphisms, and relation to disease, *Vitamins and Hormones (San Diego, CA, United States)* **2008**, *79*, 375-392.

[270] Vinukonda, G. Plasma homocysteine and methylenetetrahydrofolate reductase gene polymorphism in human health and disease: an update, *International Journal of Human Genetics* **2008**, *8*, 171-179.

[271] McCullough, J. L.; Maren, T. H. Inhibition of dihydropteroate synthetase from Escherichia coli by sulfones ad sulfonamides, *Antimicrobial Agents and Chemotherapy* **1973**, *3*, 665-9.

[272] De Benedetti, P. G.; Iarossi, D.; Menziani, C.; Caiolfa, V.; Frassineti, C.; Cennamo, C. Quantitative structure-activity analysis in dihydropteroate synthase inhibition of sulfones. Comparison with sulfanilamides, *Journal of Medicinal Chemistry* **1987**, *30*, 459-64.

[273] Zhang, Y.; Meshnick, S. R. Inhibition of Plasmodium falciparum dihydropteroate synthetase and growth in vitro by sulfa drugs, *Antimicrobial Agents and Chemotherapy* **1991**, *35*, 267-71.

[274] Chio, L.-C.; Bolyard, L. A.; Nasr, M.; Queener, S. F. Identification of a class of sulfonamides highly active against dihydropteroate synthase from Toxoplasma gondii, Pneumocystis carinii, and Mycobacterium avium, *Antimicrobial Agents and Chemotherapy* **1996**, *40*, 727-33.

[275] Prabhu, V.; Lui, H.; King, J. Arabidopsis dihydropteroate synthase: general properties and inhibition by reaction product and sulfonamides, *Phytochemistry* **1997**, *45*, 23-27.

[276] Prabhu, V.; Chatson, K. B.; Lui, H.; Abrams, G. D.; King, J. Effects of sulfanilamide and methotrexate on 13C fluxes through the glycine decarboxylase/serine hydroxymethyltransferase enzyme system in Arabidopsis, *Plant Physiology* **1998**, *116*, 137-144.

[277] Suling, W. J.; Seitz, L. E.; Reynolds, R. C.; Barrow, W. W. New Mycobacterium avium antifolate shows synergistic effect when used in combination with dihydropteroate synthase inhibitors, *Antimicrobial Agents and Chemotherapy* **2005**, *49*, 4801-4803.

[278] Nzila, A. Inhibitors of de novo folate enzymes in Plasmodium falciparum, *Drug discovery today* **2006**, *11*, 939-944.

[279] Singh, D.; Pandey, R. K. QSAR study of inhibitors of enzyme dihydropteroate synthetase, *Organic Chemistry (Rajkot, India)* **2008**, *4*, 86-90.

[280] *statistics taken from the PDB website*, http://www.pdb.org/pdb/home/home.do

[281] Development Core Team, R. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing **2009**, http://www.R-project.org

[282] Orengo, C. A. CORA-topological fingerprints for protein structural families, *Protein Science* **1999**, *8*, 699-715.

[283] Gunther, J.; Bergner, A.; Hendlich, M.; Klebe, G. Utilising Structural Knowledge in Drug Design Strategies: Applications Using Relibase, *Journal of Molecular Biology* **2003**, *326*, 621-636.

[284] Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: Design and Development of a Database for Comprehensive Analysis of Protein-Ligand Interactions, *Journal of Molecular Biology* **2003**, *326*, 607-620.

[285] Hendlich, M.; Rippmann, F.; Barnickel, G.; Hemm, K.; Aberer, K. RELIBase - an object-oriented comprehensive receptor-ligand database, *Folding & Design* **1996**, *1*, S30.

[286] Hardcastle, I. R.; Arris, C. E.; Bentley, J.; Boyle, F. T.; Chen, Y.; Curtin, N. J.; Endicott, J. A.; Gibson, A. E.; Golding, B. T.; Griffin, R. J.; Jewsbury, P.; Menyerol, J.; Mesguiche, V.; Newell, D. R.; Noble, M. E. M.; Pratt, D. J.; Wang, L.-Z.; Whitfield, H. J. N2-Substituted O6-Cyclohexylmethylguanine Derivatives: Potent Inhibitors of Cyclin-Dependent Kinases 1 and 2, *Journal of Medicinal Chemistry* **2004**, *47*, 3710-3722.

[287] Liu, M.; Choi, S.; Cuny, G. D.; Ding, K.; Dobson, B. C.; Glicksman, M. A.; Auerbach, K.; Stein, R. L. Kinetic Studies of Cdk5/p25 Kinase: Phosphorylation of Tau and Complex Inhibition by Two Prototype Inhibitors, *Biochemistry* **2008**, *47*, 8367-8377.

[288] Abrahamian, E.; Fox, P. C.; Nrum, L.; Christensen, I. T.; Thogersen, H.; Clark, R. D. Efficient Generation, Storage, and Manipulation of Fully Flexible Pharmacophore Multiplets and Their Use in 3-D Similarity Searching, *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 458-468.

[289] Bhattacharya, A.; Wunderlich, Z.; Monleon, D.; Tejero, R.; Montelione Gaetano, T. Assessing model accuracy using the homology modeling automatically software, *Proteins* **2008**, *70*, 105-18.

[290] Dalton, J. A. R.; Jackson, R. M. An evaluation of automated homology modelling methods at low target-template sequence similarity, *Bioinformatics* **2007**, *23*, 1901-1908.

[291] Kiefer, F.; Arnold, K.; Kunzli, M.; Bordoli, L.; Schwede, T. The SWISS-MODEL Repository and associated resources, *Nucleic Acids Research* **2009**, *37*, D387-92.

[292] Moglich, A.; Weinfurtner, D.; Maurer, T.; Gronwald, W.; Kalbitzer Hans, R. A restraint molecular dynamics and simulated annealing approach for protein homology modeling utilizing mean angles, *BMC Bioinformatics* **2005**, *6*, 91.

[293] Takeda-Shitaka, M.; Terashi, G.; Chiba, C.; Takaya, D.; Umeyama, H. FAMS Complex: a fully automated homology modeling system for protein complex structures, *Medicinal Chemistry* **2006**, *2*, 191-201.

[294] Bujnicki, J. M.; Elofsson, A.; Fischer, D.; Rychlewski, L. Livebench-2: large-scale automated evaluation of protein structure prediction servers, *Proteins: Structure, Function, and Genetics* **2002**, 184-191.

[295] Zhang, Y.; Skolnick, J. Automated structure prediction of weakly homologous proteins on a genomic scale, *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 7594-7599.

[296] Battey, J. N. D.; Kopp, J.; Bordoli, L.; Read, R. J.; Clarke, N. D.; Schwede, T. Automated server predictions in CASP7, *Proteins: Structure, Function, and Bioinformatics* **2007**, *69*, 68-82.

[297] Arve, L. *Synthese und Konformationsanalyse von Biphenomycin-Analoga*; Technical University of Dortmund; 2006.

[298] Bisek, N.; Wetzel, S.; Arndt, H.-D.; Waldmann, H. Synthesis and conformational analysis of stevastelin C3 analogues and their activity against the dual-specific vaccina H1-related phosphatase, *Chemistry--A European Journal* **2008**, *14*, 8847-8860.

[299] Bisek, N. *Synthese und Konformationsanalyse von Stevastelin-C3-Analoga als Phosphataseinhibitoren*; Technical University of Dortmund; 2008.

[300] Koehn, M.; Gutierrez-Rodriguez, M.; Jonkheijm, P.; Wetzel, S.; Wacker, R.; Schroeder, H.; Prinz, H.; Niemeyer, C. M.; Breinbauer, R.; Szedlacsek, S. E.; Waldmann, H. A microarray strategy for mapping the substrate specificity of protein tyrosine phosphatase, *Angewandte Chemie, International Edition* **2007**, *46*, 7700-7703.

[301] Guo, Z.; Wu, Y.-W.; Tan, K.-T.; Bon, R. S.; Guiu-Rozas, E.; Delon, C.; Nguyen, U. T.; Wetzel, S.; Arndt, S.; Goody, R. S.; Blankenfeldt, W.; Alexandrov, K.; Waldmann, H. Development of selective RabGGTase inhibitors and crystal structure of a RabGGTase-inhibitor complex, *Angewandte Chemie, International Edition* **2008**, *47*, 3747-3750.

[302] Tan, K.-T.; Guiu-Rozas, E.; Bon, R. S.; Guo, Z.; Delon, C.; Wetzel, S.; Arndt, S.; Alexandrov, K.; Waldmann, H.; Goody, R. S.; Wu, Y.-W.; Blankenfeldt, W. Design, Synthesis and Characterization of Peptide-Based RabGGTase Inhibitors, *Journal of the American Chemical Society* **2009**.

[303] Koch, M. A. *Protein- und Naturstoffstruktur als Leitprinzipien für die Entwicklung von Verbindungsbibliotheken*; University of Dortmund; 2005.

[304] Dekker, F. J.; Rocks, O.; Vartak, N.; Balamurugan, R.; Menninger, S.; Wetzel, S.; Renner, S.; Gerauer, M.; Hedberg, C.; Kramer, J. W.; Coates, G. J.; Brunsveld, L.; Bastiaens, P.; Waldmann, H. Small molecule inhibition of depalmitoylation reverts unregulated Ras signaling, *Nature (London, United Kingdom)* **2009**.

[305] Triola, G.; Wetzel, S.; Ellinger, B.; Koch, M. A.; Huebel, K.; Rauh, D.; Waldmann, H. ATP competitive inhibitors of -alanine--alanine ligase based on protein kinase inhibitor scaffolds, *Bioorganic & Medicinal Chemistry* **2009**, *17*, 1079-1087.

[306] Chemical Abstract Services **2009**, http://www.cas.org/expertise/cascontent/registry/index.html

[307] Gnerre, C.; Catto, M.; Leonetti, F.; Weber, P.; Carrupt, P.-A.; Altomare, C.; Carotti, A.; Testa, B. Inhibition of Monoamine Oxidases by Functionalized Coumarin Derivatives: Biological Activities, QSARs, and 3D-QSARs, *Journal of Medicinal Chemistry* **2000**, *43*, 4747-4758.

[308] Bruehlmann, C.; Ooms, F.; Carrupt, P.-A.; Testa, B.; Catto, M.; Leonetti, F.; Altomare, C.; Carotti, A. Coumarins derivatives as dual inhibitors of acetylcholinesterase and monoamine oxidase, *Journal of Medicinal Chemistry* **2001**, *44*, 3195-3198.

[309] Harfenist, M.; Heuser, D. J.; Joyner, C. T.; Batchelor, J. F.; White, H. L. Selective Inhibitors of Monoamine Oxidase. 3. Structure-Activity Relationship of Tricyclics Bearing Imidazoline, Oxadiazole, or Tetrazole Groups, *Journal of Medicinal Chemistry* **1996**, *39*, 1857-63.

[310] Rooke, N.; Li, D.-J.; Li, J.; Keung, W. M. The Mitochondrial Monoamine Oxidase-Aldehyde Dehydrogenase Pathway: A Potential Site of Action of Daidzin, *Journal of Medicinal Chemistry* **2000**, *43*, 4169-4179.

[311] Gao, G.-Y.; Li, D.-J.; Keung, W. M. Synthesis of Potential Antidipsotropic Isoflavones: Inhibitors of the Mitochondrial Monoamine Oxidase-Aldehyde Dehydrogenase Pathway, *Journal of Medicinal Chemistry* **2001**, *44*, 3320-3328.

8 Glossary

11β-HSD	11β-Hydroxysteroid dehydrogenase (11β-HSD) is an enzyme that exists in the two subtypes 1 and 2. Whereas 11β-HSD1 reduces cortisone to the active hormone cortisol, 11β-HSD2 catalyzes the reverese reaction. By regulation of cortisol levels, both 11β-HSD enzymes influence the activity of glucocorticoid and mineralcorticoid receptors. 11β-HSD1 is considered a drug target in the treatment of obesity, metabolic syndrome and diabetes type 2.
AC_{50}	The 'active concentration 50' (AC_{50}) is the compound concentration at which the half-maximal effect of a corresponding control value is observed. It replaces the 'inhibitory concentration 50' (IC_{50}) in cases where, for example, inhibition and activation are measured in the same experiment.
AChE	Acetylcholinesterase (AChE) is an enzyme that cleaves the neutransmitter acetylcholine, thereby terminating the signal transmission at cholinergic synapses. AChE is a drug target, for example in Alzheimer's disease.
ADP	Adenosine diphosphate (ADP) is a nucleotide that contains a pyrophosphate unit, i.e. two joined phosphates. ADP and its phosphorylated form, ATP, play an important role in energy storage and transport in living systems.
aSMase	Acid sphingomyelinase (see sphingomyelinase)
ATCC	The American Type Culture Collection (ATCC) is a non-profit organization offering access to standardized biological materials like cell lines or microorganisms.
ATP	Adenosine-triphosphate (ATP) is a nucleotide that contains three joined phosphates. ATP and its de-phosphorylated form, ADP, play an important role in energy storage and transport in living systems.
Batik	Batik is a Java-based toolkit that offers functionality for the processing of SVG images. It is available free of charge under an open source license from http://xmlgraphics.apache.org/batik/index.html.
BIOS	Biology-inspired synthesis (BIOS) aims at the generation of small, focused compound collections that are generated from biologically pre-validates structures, for instance natural products. BIOS also integrates bio- and cheminformatics strategies as means to enrich the library with biological relevance for the protein target of interest.

Brachiation	Brachiation describes the aboreal locomotion, i.e. the swinging from branch to branch, of gibbons in botanical trees. In this work it was adapted to denote the movement from large, more complex scaffolds towards smaller, less complex structures along the branches of the scaffold tree while retaining similar, yet varied bioactivity.
BSA	Bovine serum albumin (BSA) is a serum albumin that is often used in biochemical assays as an unspecific binding protein.
CA	Carbonic anhydrase catalyzes the rapid conversion of carbon dioxide to carbonic acid, for instance in the kidney.
CAS registry	CAS database holding more than 46 million organic and inorganic compounds as well as more than 60 million sequences as of May 2009.[306]
CAS	Chemical Abstracts Services, a branch of the American Chemical Society.
CATH	The CATH database provides a structure-based hierarchical classification of proteins, similarly to SCOP. In contrast to SCOP, the protein structures are assigned to their hierarchy class by a combination of automated and manual methods.
Cdc25A	The dual-specificity phosphatase Cdc25A is involved in regulation of the cell cycle. Therefore, it is considered a drug target in anti-cancer therapy.
CE	The Combinatorial Extensions algorithm aligns 3D protein structures with one another.
ChemGPS	ChemGPS is a property-based approach to charting of and navigation through chemical space developed by Oprea *et al*. A 3D diagram representation of chemical space is generated by interpolation of the position of each compound (= dot) according to a reference set. The ChemGPS reference set was extended to natural product chemical space by Larrson *et al*.
COM	The centre of mass (COM) is the point where the individual masses of a particle swarm behave as one concentrated mass.
CSA	The Catalytic Site Atlas (CSA) is a database that contains the catalytic residues of enzymes annotated from more than 1,000 literature reference and by sequence alignments.
CV	Column volume
Dali	The Dali algorithm structurally aligns protein structures by their tertiary structure and determines the structural similarity.

DHPS	Dihydropteroate synthetase (DHPS) belongs to the folate pathway and converts dihydropteroate diphosphate and 4-Aminobenzoic acid into dihydropteroic acid, a folate precursor. It is found only in bacteria but not in humans and inhibited by sulfonamides rendering it an antibiotic drug target.
DMSO	Dimethylsulfoxide (DMSO) is a standard solvent in chemistry.
DNP	The Dictionary of Natural Products (DNP) is one of the most comprehensive archives of natural products and analogues. It is compiled from literature and is currently marketed as an online, offline and a cheminformatics version by Chapman & Hall / CRC Informa.
DOS	Diversity-oriented synthesis (DOS) was developed by Stewart Schreiber during the late 1990s. It denotes a synthetic strategy which creates structural diversity by combinatorial synthesis during the generation of the molecular scaffolds as opposed to combinatorial synthesis strategies where a common scaffold is diversified by the attachment of different substituents at pre-defined positions. DOS generates large libraries maximising diversity and complexity.
Drugbank	Drugbank is a database that combines chemical and pharmaceutical information about drugs with detailed protein target annotation. It can be accessed free of charge at http://www.drugbank.ca/.
DSSP	The Dictionary of Protein Secondary Structure (DSSP) was developed by Kabsch and Sanders to predict secondary structure elements from protein structure files. It still represents the gold standard in secondary structure detection.
DTT	Dithiotreitol (DTT) is a reducing agent that is often used to prevent the formation of inter- and intramolecular disulfide bonds between cystein residues in proteins in solution.
EDTA	Ethylenediaminetetraaceticacid (EDTA) is a chelating agent that strongly binds di- and tricationic metal ions, thereby removing them from solution.
E-FABP	Epidermal fatty acid binding protein (E-FABP) is believed to be involved in fatty acid transport and metabolism. It has recently been described as a biomarker associated with cardio-metabolic risk factors and carotid arteriosclerosis.
EGTA	Ethyleneglycoltetraaceticacid (EDTA) is a chelating agent that strongly binds di- and tricationic metal ions, thereby removing them from solution. In contrast to EDTA, it binds calcium ions much stronger than magnesium ions.

FBDD	Fragement-based drug discovery (FBDD) was pioneered by the company Astex in the late 1990s. FBDD screens very small, low-affinity molecules with a molecular weight below 240 daltons in high concentrations in biochemical assays and by structural techniques like protein crystallography or NMR. Successful fragments are then developed to inhibitor molecules, for instance by linking or synthetic growth, guided by the structural information.
FLAP	The fingerprints for ligands and proteins (FLAP) were developed in the group of Gabriele Cruciani at the University of Perugia, Italy. They describe proteins and small molecules by pharmacophoric features on their surface, for example the hydrophilicity or the hydrogen bond donors and acceptors using the GRID probes.
FSSP	The Families of Structurally Similar Protein (FSSP) database was designed by Lisa Holm and Chris Sander. It contains structural alignments of all available protein structures precomputed by Dali. It should be noted, however, that FSSP only calculates the structural alignments between the centres of protein clusters defined by more than 70% sequence identity.
Hepes	Hepes is the acronym of 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid, a zwitterionic buffer molecule. It is often used in biochemical buffer solutions.
HMU-PC-substrate	6-Hexadecanoylamino-4-methylumbelliferylphosphorylcholine (HMU-PC) is a fluorogenic substrate for sphingomyelinase assays.
HTS	High throughput screening facilitates the testing of sets of several thousand up to several million compounds for their biological activity against a protein of interest within several day or weeks. It relies heavily on automation by specialized robots and computers as well as advanced statistical data treatment.
Java	The programming language Java was developed during the 1990's to enable platform-independent programming, i.e. facilitate the coding of programs that readily can be executed under different operating systems, such as Microsoft Windows, Linux or MacOS.
JDBC	The Java Database Connector (JDBC) provides a standardized framework to access databases from the programming language Java. The Java commands are translated by a JDBC driver for the individual database system that should be accessed. JDBC drivers are available for all major database systems.

KCl	Sodium chloride
LBE	Ligand-binding efficiency was introduced to better compare binding affinities of protein ligands spanning a wide size range. It is calculated by division of the free energy or, in most cases, the pIC_{50} (= - log IC_{50}) by the number of heavy atoms in the molecule. Compounds with LBEs above 0.4 are usually selected for further development.
Lck	Leukocyte-specific protein tyrosine kinase (Lck) is a protein tyrosine kinase and a member of the Src kinase family.
LDH	Lactate dehydrogenase (LDH) is a metabolic enzyme that catalyzes the reduction of pyruvate to lactate under consumption of NADH.
M6P	Mannose 6-posphate
M6P-IGF2R	Mannose 6-phosphate/insulin-like growth factor II receptor is a protein that recruits insulin-like growth factor 2 (IGF2) at the cell membrane followed by internalization in vesicles. It shuttles IGF2 to endosomes, thereby terminating IGF2-mediated signalling.
MAO	The monoamine oxidases are an enzymes that utilize a confactor (FAD) to oxidatively de-aminate a wide range of amines. The two subtypes, MAO A and B differ in their natural substrates. Mao A oxidizes 5-hydroxytryptamine (serotonin) while MAO B converts benzylamine and 2-phenyethylamine. Both enzymes are targets in anti-depressive drug development and in the therapy of Alzheimer's and Parkinson's diseases.
Me-too drug	Me-too drugs are analogue-based drugs that are developed on the basis of a known drug, often from a competitor company. In many cases the compounds share structural motifs or pharmacophores.
MTHFR	Methylenetetrahydrofolate reductase (MTHFR) reduces 5,10-methylenetetrahydrofolate to 5,10-methyltetrahydrofolate in the folate pathway.
MTHMP DH	Methylenetetrahydromethanopterin dehydrogenase (MTHMP DH) is an oxidoreductase catalyzing the co-enzyme F420-mediated oxidation of methylenetetrahydromethanopterin. It is part of the folate biosynthesis pathway.
NADH	Nicotinamide adenine dinucleotide (NADH) is a co-factor found in two oxidation states: the reduced form NAD^+ and the oxidized form NADH. The molecule serves as electron acceptor or donor during intracellular redox reactions.
NMR	Nuclear magnetic resonance (NMR) is a standard technique to determine protein structures by analysis of interference patterns of signals

	generated by radio-frequency induced shifts of the magnet moment of atom cores.
NP40	NP-40 or tergitol-type NP-40 describes the molecule nonyl phenoxylpolyethoxylethanol, a detergent that is often used in biochemical assays to prevent compound aggregation at higher concentrations.
nSMase	Neutral sphingomyelinase (see sphingomyelinase)
OptiSim	The OptiSim clustering algorithm was designed to generate a diverse set of cluster centres from a collection of entities.
PBS	Phosphate buffered saline (PBS) is a buffer solution containing mainly sodium chloride and sodium phosphate.
PCA	Principal component analysis (PCA) is a mathematical transformation that coverts the set of basis vectors of an n-dimensional vector space into a set of basis vectors of an m-dimensional vector space whereas m < n.
PDB	The Protein Data Bank (PDB) is the world's most comprehensive resource of protein structures. As of June 2009 it comprises of more than 58,000 protein structures.
PEP	Phosphoenolpyruvate (PEP) is a substrate of pyruvate kinase that de-phosphorylates it to pyruvate.
Piccolo	The Piccolo toolkit is Java-based and facilitates the generation of structured 2D graphics, e.g. graphs. It is available free of charge under an open source license from http://www.cs.umd.edu/hcil/jazz/.
PipelinePilot	The workflow tool PipelinePilot was developed by SciTegic, now a part of Accelrys. It offers a broad range of cheminformatics functionality in an easy-to-use format.
PK	Pyruvate kinase is involved in glycolysis where it catalyzes the de-phosphorylation of phosphoenolpyruvate to pyruvate while converting one molecule of ADP to ATP, i.e. storing the release energy in a bioavailable form. Pyruvate kinase was described as a drug target in Malaria therapy.
PSI-BLAST	The position specific iterative BLAST algorithm is used to align protein sequences and determine their similarity. It is an improved version of the BLAST algorithm that is more senstitive to more distantly related protein sequences.
PSSC	Protein Structure Similarity Clustering (PSSC) is a concept that builds on structural complementarity between scaffold in proteins and compounds. It hypothesizes that proteins with a similar subfold in the active site bind similar ligands, thereby defining target clusters addressable by compound

	libraries designed around the scaffold of a known inhibitor of one of the cluster member proteins.
PubChem	The PubChem database was created withing the Molecular Libraries Roadmap Initiative of the National Institutes of Health (NIH) in the USA. It is supposed to serve biomedical research by providing publicly available structure-related bioactivity data for free. As of July 2009, it contained more than 37 million unique chemical structures and more than 1,500 bioassays.
R5P	Ribose-5-phosphate (R5P) is a phosphorylated pentose sugar that activates the pyruvate kinase from *bacillus stearothermophilus*, presumably by an allosteric mechanism.
RECAP	The Retrosynthetic Combinatorial Analysis Procedure (RECAP) created a set of rules to dissect molecules at functional groups resembling retrosynthetic dissections used by chemists.
Relibase	Relibase is a commercial database that contains the curated protein structures from the PDB. It offers additional annotation of binding sites, ligands and other properties, together with graphical tools for the detailed analysis of large sets of protein structures.
RMSD	The root mean square deviation (RMSD) is a statistical measure for the similarity of protein structures. It is calculated as a statistical average over all the distances between the corresponding atoms in two protein structures. The smaller the RMSD is, the more similar both structures are; RMSD values of smaller than 1 indicate identical structures.
Rule-of-Five	The Rule-of-Five is an empirical set of parameters to predict compounds which are likely to be orally available drugs. It was derived from known orally available drugs by Christopher Lipinski in 1997. In short, the suitable compounds have a molecular weight below 500, an octanol-water partition coefficient (log P) of less than 5 and contain no more than five hydrogen bond donors and no more than 10 hydrogen bond acceptors.
SAR	Structure-activity relationships describe the influence of structural variation of residues attached to a common core structure on the biological activity of the resulting compounds. It is important to note that for SAR generation active and inactive molecules are need. A systematic variation of the different residues is favourable.
Scaffold Hunter	Scaffold Hunter is a Java-based computer program that offers a chemically intuitive, automated, and interactive visualization of chemical

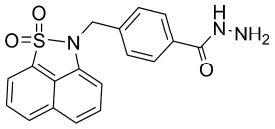
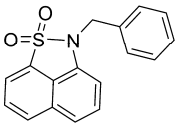
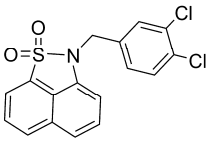
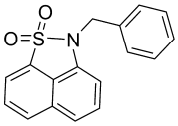
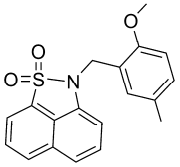
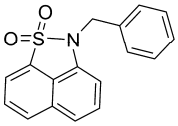
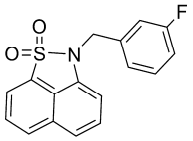
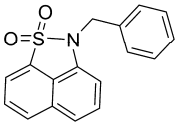
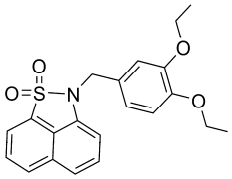
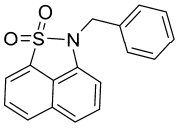
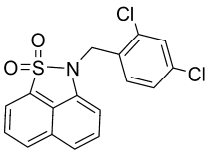
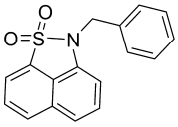
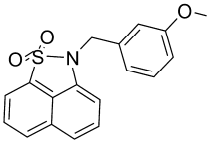
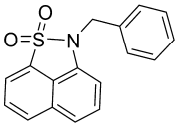
	space by the means of scaffold trees. It facilitates analyis and mining of structure-related bioactivity data by eductated non-expert users, e.g. chemists and biologists.
Scaffold tree	The scaffold tree evolved from the Structural Classification of Natural Products (SCONP) by implementation of a new set of rules guiding the scaffold tree construction process. One major difference is that scaffold trees comprise also virtual scaffolds that are not pesent in molecules in the data set whereas SCONP does not allow virtual scaffolds in the tree-like diagram.
SCONP	The Structural Classification of Natural Products (SCONP) was developed as a means for charting natural product chemical space. It generated a tree-like diagram of a scaffold hierarchy in which exclusively scaffolds occurring in natural products were allowed.
SCOP	The Structural Classification of Proteins (SCOP) database hierarchically classifies proteins with regards to their structural relationship using manual assignment by domain experts and automatic pre-categorization.
SD File	The SD File format was developed by MDL. It can contain records for multiple molecules each of which is described by its atom and bond definitions as well as additional information. Since the definition of the SD File format is publicly available it became a standard format for exchange of molecular information that is supported by almost a wide range of applications.
SDS-PAGE	SDS-PAGE describes the sodium dodecyl sulphate (SDS) polyacrylamide gel electrophoresis (PAGE), a technique for the analysis of proteins. During the analysis, the proteins in the test sample are first denatured by mixing with SDS, a detergent. In the next step, the protein chains are separated in a gel electrophoresis. This technique is based on the different speed of moving of protein chains in an electric field in a gel depending on their size.
SGE	Sun Grid Engine (SGE) is a program for the submission, distribution and administration of computational jobs over a linux cluster. It is maintained by the company SUN and available free of charge.
SH2 domain	The Src Homology 2 (SH2) domain is conserved in many proteins involved in intracellular signalling. It facilitates protein-protein binding by recognition of the phosphorylation state of tyrosine residues, that is, only binding partnerswith one or more phosphorylated tyrosine residues are

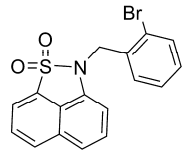
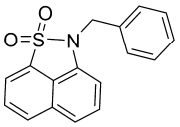
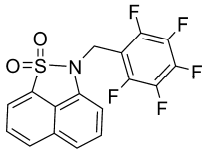
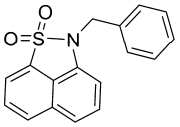
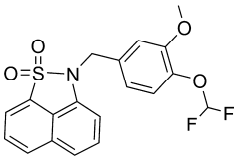
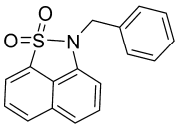
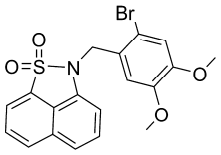
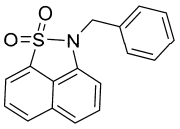
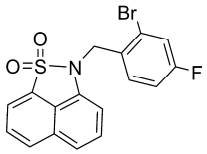
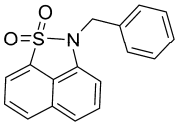
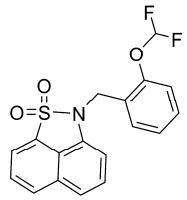
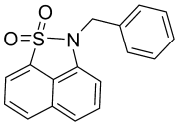
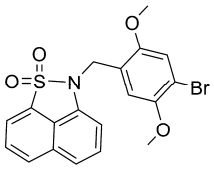
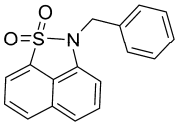
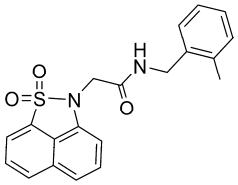
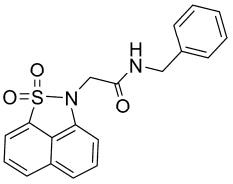
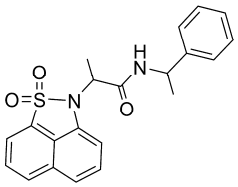
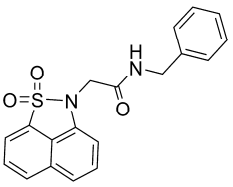
	bound. This binding event is often followed by a cascade of protein-protein interactions resulting in many different biological responses.
SM	Sphingomyelin, a sphingoplipid found, for example, in the myeline sheaths around the nerve axons. It is a natural substrate of sphingomyelinase, which cleaves it into phosphorylcholine and ceramide.
SMILES	The Simplified Molecular Input Line Entry System (SMILES) was developed by Dave Weininger in 1988 to encode the structure of a molecule in a text string. However, a give molecule can be represented by many different SMILES strings which led to the development of canonical SMILES where each molecule is represented by exactly one unique SMILES string. One drawback is that canonical SMILES are only unique within the software that generated them but not across different software packages.
Sphingomyelinase	The sphingomyelinases are hydrolases involved in sphingolipid metabolism. The neutral isoenzyme emerged as a drug target in platelet aggregating factor (PAF) induced lung edema, a major health problem in intensive care patients.
SQL	The Structured Query Language (SQL) is a standard computer language for database operations including data input and retrieval. It is widely supported by current database systems and forms a *de-facto* standard.
SSE	Secondary structure elements (SSEs) are small, stable structural units, for example, α-helices or β-sheets that are formed by amino acid sequences due to intramolecular forces.
STAT	The Signal Transducers and Activator of Transcription (STAT) proteins are transcriptionfactors regulation cell growth and differentiation. They are usually activated by phosphorylation by kinases upon which they dimerize. The dimmers translocate to the nucleus where they induce gene transcription.
SVG	The graphics format Scalable Vector Graphics (SVG) is an XML-based format that stores image information in a vectorized format, that is, as lines, circles etc. rather than as pixels. SVG is text based, therefore, the resulting graphics are rather small compared to bitmap images. The vectorized nature retains high quality graphical representation during scaling since no interpolation of pixels is needed.
TByte	One Tera Byte (TByte) comprises of 1,000 Gigabytes or 1,000,000,000,000 Bytes.

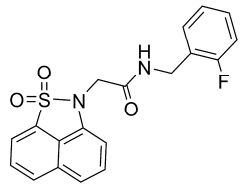
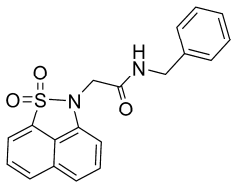
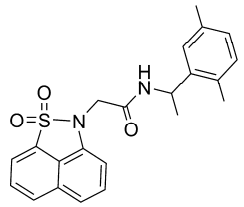
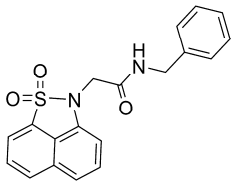
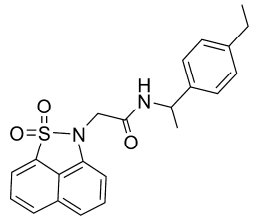
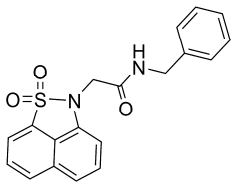
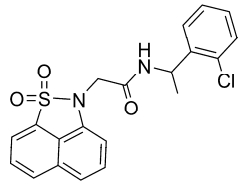
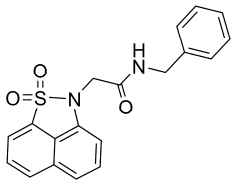
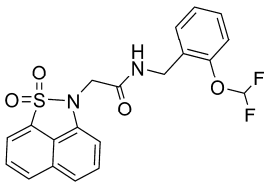
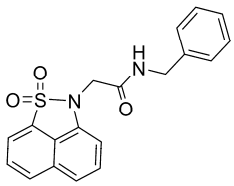
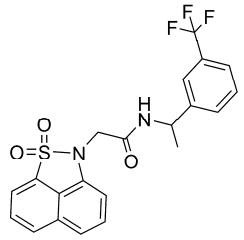
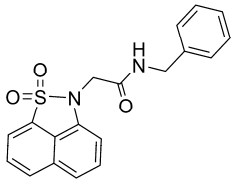
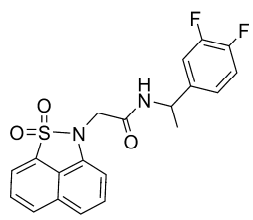
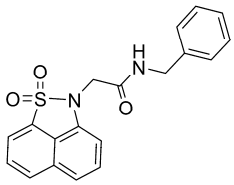
Tris	Tris is the abbreviation for tris(hydroxymethyl)aminomethane with a pKa of 8.06 that is used as a component in buffers.
Triton X-100	Triton X-100 is a non-ionic sufectant that is often used in biochemical assays to prevent compound aggregation at higher concentrations.
Tween 20	Tween 20 or Polysorbate 20 is a polysorbate surfactant that is often used in biochemical assays to prevent compound aggregation at higher concentrations.
UniProt	The Universal Protein database (UniProt), on of the most comprehensive protein resources on the worldwide web, provides information, e.g. sequence, associated genes, links to data stored in other databases, for more than 470,000 proteins as of July 2009. Its identifier is often used to provide a common reference to a particular protein especially if protein data from multiple sources is compared.
VAST	The vector aligment search tool aligns protein structures. It uses vector representations of secondary structure elements and aligns the resulting vector fields.
WOMBAT	The World of Molecular Bioactivity (WOMBAT) database is a compilation of small molecule structures together with their bioactivity and protein targets extracted from literature.
XLFit	XLFit is a plugin for Excel that offers extensive curve-fitting capabilities. It is available from IDBS (http://www.idbs.com/decision/xlfit/).
XML	The data format extensible markup language (XML) is a text-based format for structured data. It operates with so-called ‘tags’, each of which defines an information category and is included in angle brackets: <tag>. Today, XML is used widely in data storage and exchange because of its flexibility.
Z-Score	The Z-Score is a measure for the statistical significance of structural alignments. Alignments of highly similar structures typically yield values around 25-36 where as values smaller than 2 are considered insignificant. As the Z-Score is a statistical value, it correlates strongly with the number of aligned residues, i.e. the more residues are aligned, the higher the resulting Z-Score.

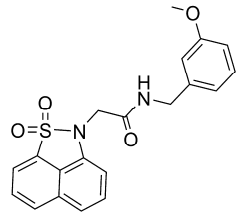
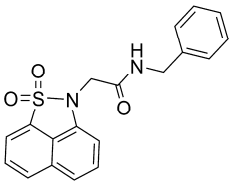
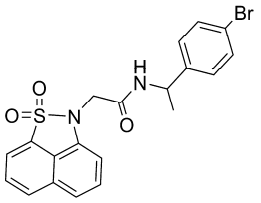
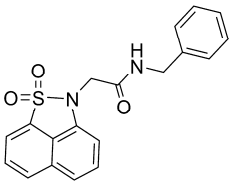
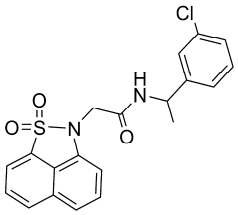
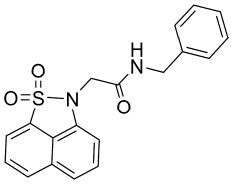
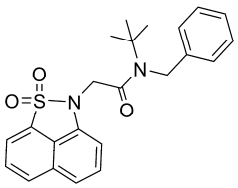
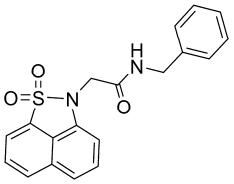
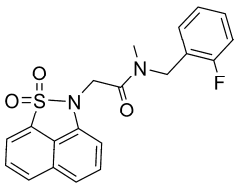
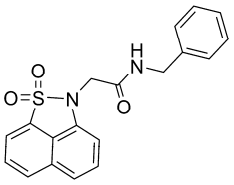
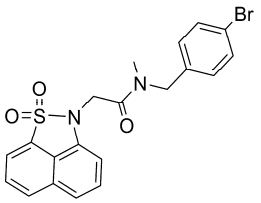
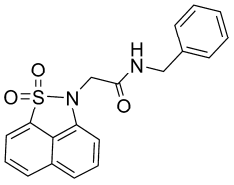
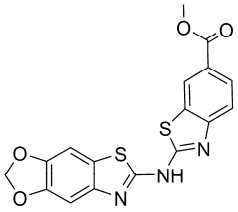
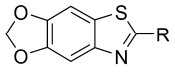
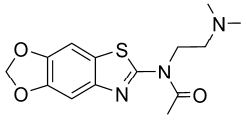
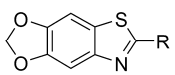
Attachments

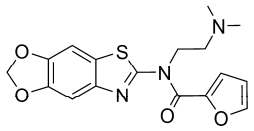
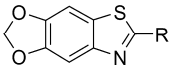
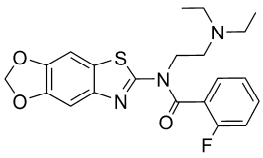
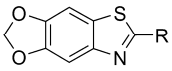
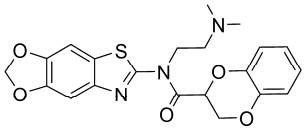
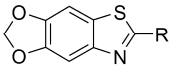
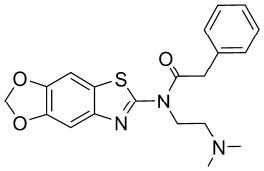
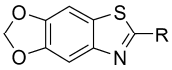
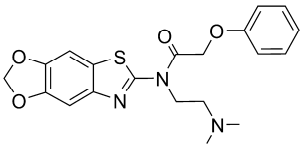
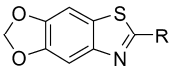
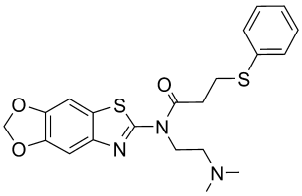
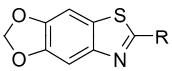
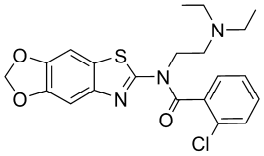
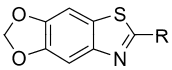
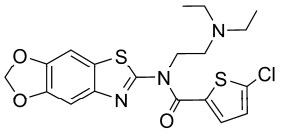
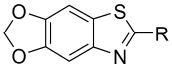
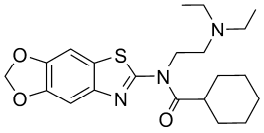
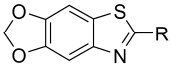
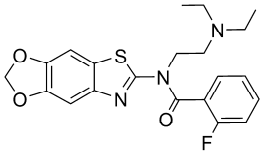
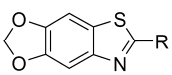
Attachment 1: Table of the 107-membered library of potential modulatory chemotypes of pyruvate kinase. The table gives the reference that is used throughout this work, the structure, and the virtual scaffold as well as the supplier code, the supplier and the PubChem compound id for each compound.

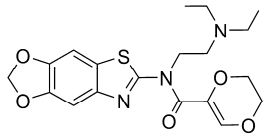
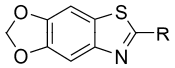
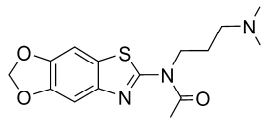
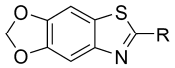
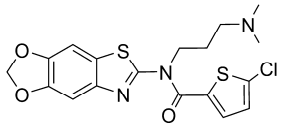
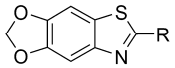
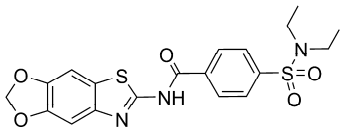
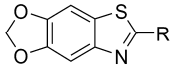
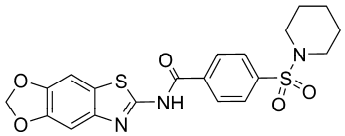
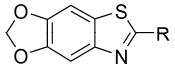
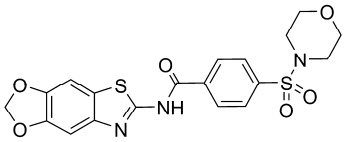
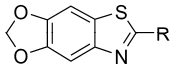
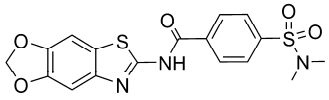
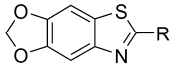
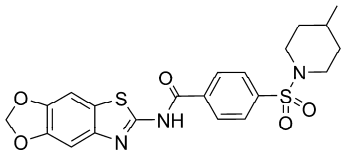
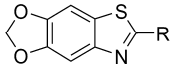
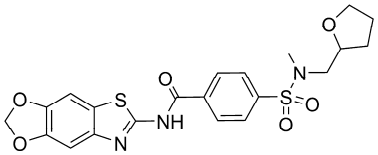
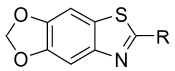
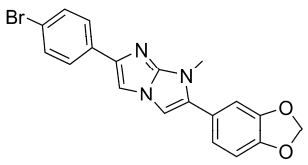
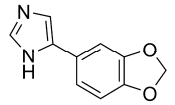
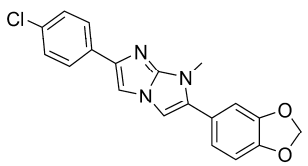
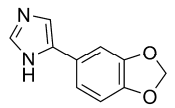
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-1			kbsenon-0012203	Aurora Fine Chemicals	57571219
PKL-2			kcd-124573	Aurora Fine Chemicals	57571220
PKL-3			kcheb-067808	Aurora Fine Chemicals	57571221
PKL-4			kcheb-070443	Aurora Fine Chemicals	57571222
PKL-5			kcheb-072091	Aurora Fine Chemicals	57571223
PKL-6			kcheb-079810	Aurora Fine Chemicals	57571224
PKL-7			kcheb-100864	Aurora Fine Chemicals	57571225

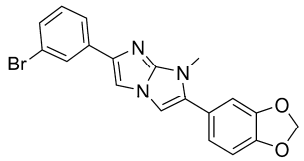
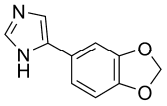
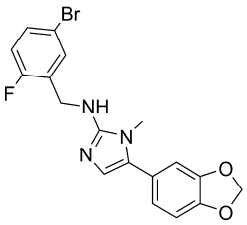
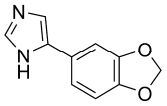
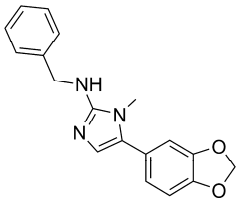
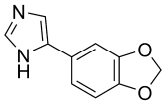
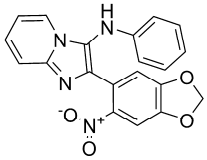
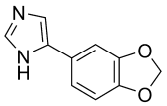
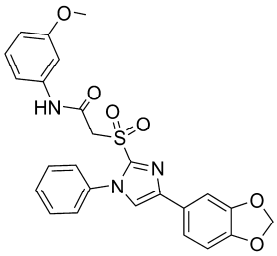
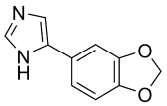
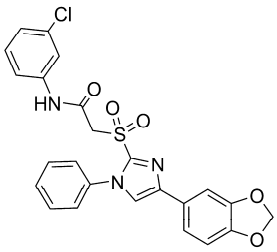
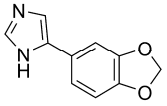
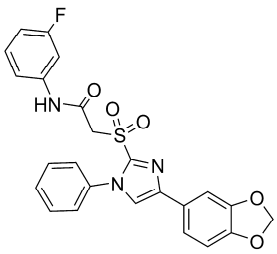
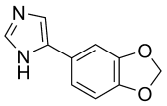
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-8			kuk-303437	Aurora Fine Chemicals	57571226
PKL-9			kuk-303461	Aurora Fine Chemicals	57571227
PKL-10			kuk-539878	Aurora Fine Chemicals	57571228
PKL-11			kuk-544394	Aurora Fine Chemicals	57571229
PKL-12			kuk-732953	Aurora Fine Chemicals	57571230
PKL-13			kuk-744159	Aurora Fine Chemicals	57571231
PKL-14			kuk-745624	Aurora Fine Chemicals	57571232
PKL-15			kchi-145021	Aurora Fine Chemicals	57571233
PKL-16			ken-252537	Aurora Fine Chemicals	57571234

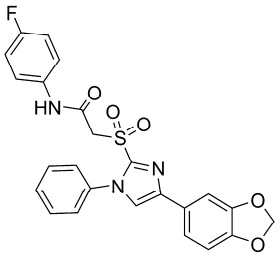
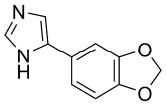
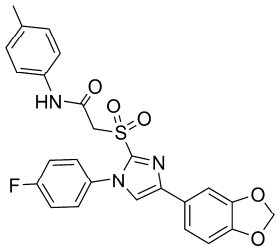
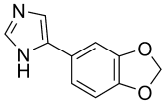
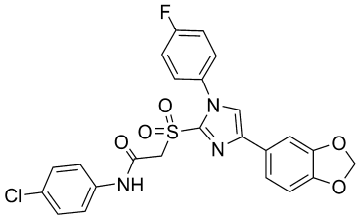
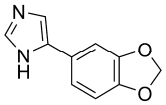
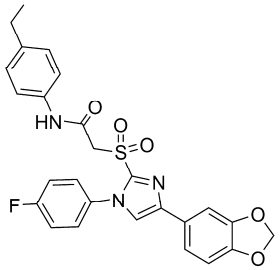
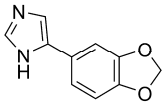
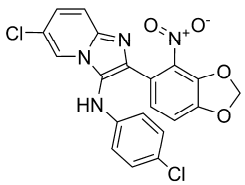
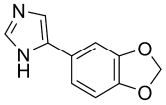
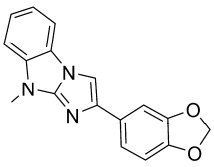
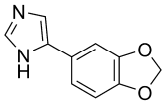
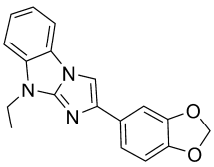
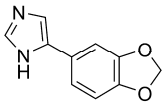
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-17			ken-255165	Aurora Fine Chemicals	57571235
PKL-18			ken-279479	Aurora Fine Chemicals	57571236
PKL-19			ken-281857	Aurora Fine Chemicals	57571237
PKL-20			ken-281858	Aurora Fine Chemicals	57571238
PKL-21			ken-289492	Aurora Fine Chemicals	57571239
PKL-22			ken-290647	Aurora Fine Chemicals	57571240
PKL-23			ken-290648	Aurora Fine Chemicals	57571241

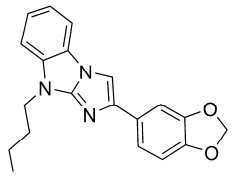
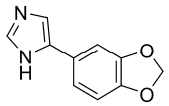
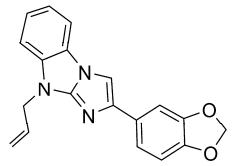
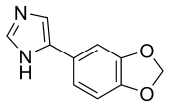
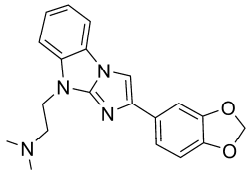
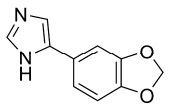
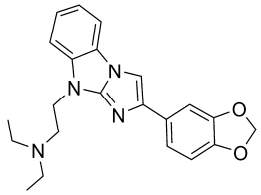
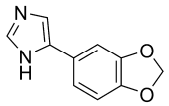
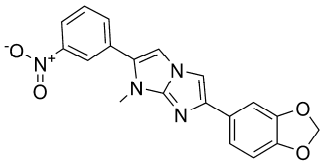
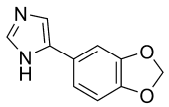
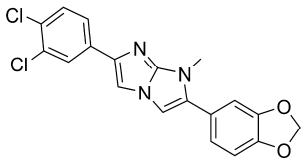
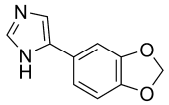
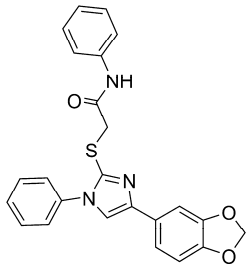
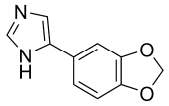
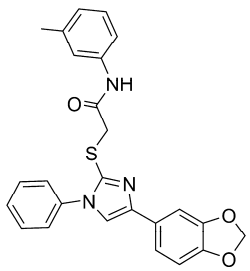
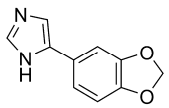
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-24			ken-333311	Aurora Fine Chemicals	57571242
PKL-25			ken-350671	Aurora Fine Chemicals	57571243
PKL-26			ken-350759	Aurora Fine Chemicals	57571244
PKL-27			ken-464194	Aurora Fine Chemicals	57571245
PKL-28			kuk-867448	Aurora Fine Chemicals	57571246
PKL-29			kuk-303440	Aurora Fine Chemicals	57571247
PKL-30			C505-0827	ChemDiv	57571141
PKL-31			E676-1525	ChemDiv	57571148

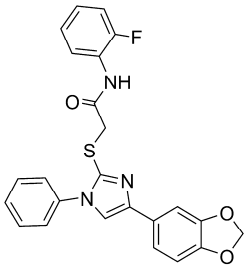
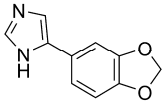
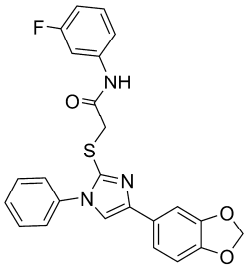
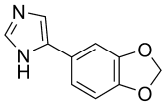
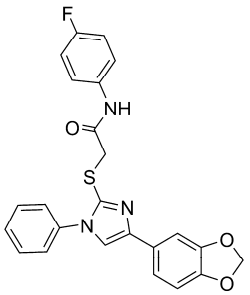
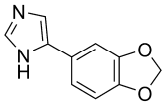
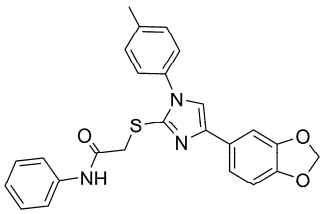
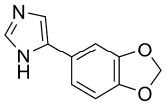
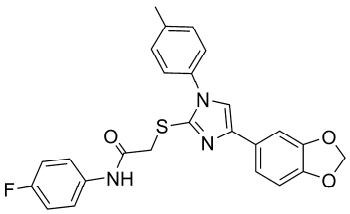
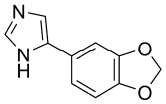
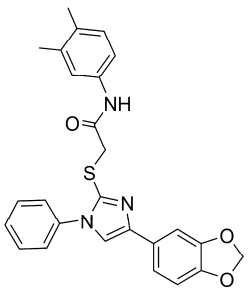
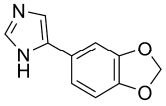
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-32			E676-1559	ChemDiv	57571149
PKL-33			E676-1574	ChemDiv	57571150
PKL-34			E676-1627	ChemDiv	57571151
PKL-35			E676-1637	ChemDiv	57571152
PKL-36			E676-1644	ChemDiv	57571153
PKL-37			E676-1645	ChemDiv	57571154
PKL-38			E676-4324	ChemDiv	57571155
PKL-39			E676-4359	ChemDiv	57571156
PKL-40			E676-4364	ChemDiv	57571157
PKL-41			E676-4404	ChemDiv	57571158

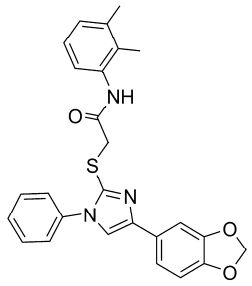
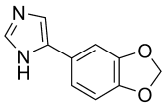
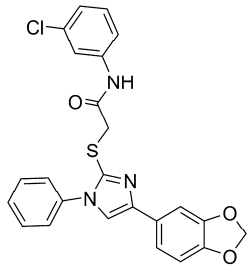
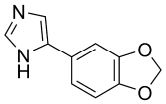
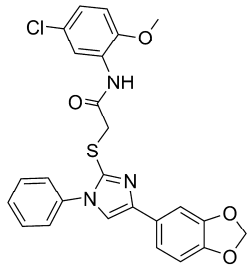
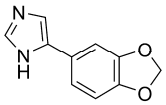
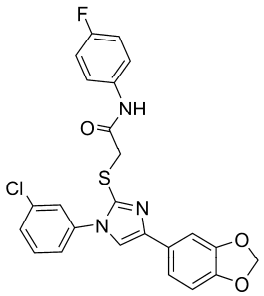
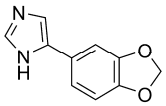
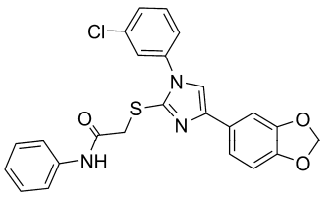
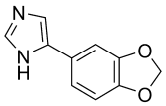
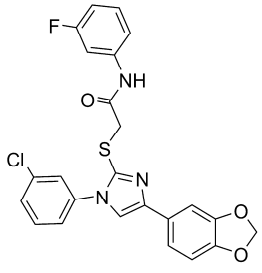
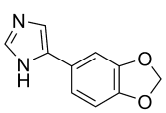
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-42			E676-4435	ChemDiv	57571159
PKL-43			E677-1525	ChemDiv	57571160
PKL-44			E677-1565	ChemDiv	57571161
PKL-45			G786-0905	ChemDiv	57571164
PKL-46			G786-0908	ChemDiv	57571165
PKL-47			G786-0909	ChemDiv	57571166
PKL-48			G786-0911	ChemDiv	57571167
PKL-49			G786-0916	ChemDiv	57571168
PKL-50			G786-0919	ChemDiv	57571169
PKL-51			7756-0336	ChemDiv	57571205
PKL-52			7756-0337	ChemDiv	57571206

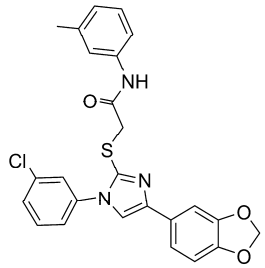
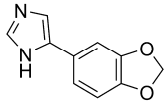
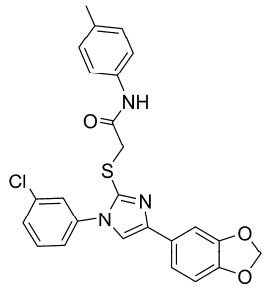
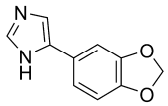
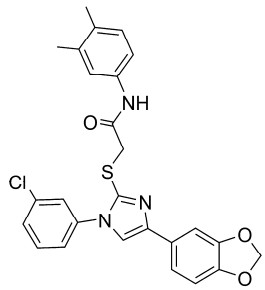
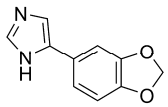
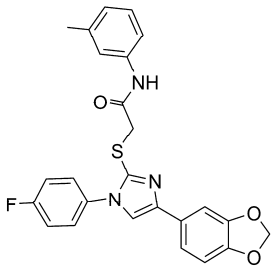
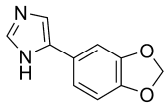
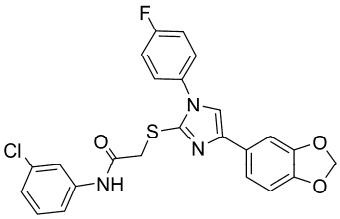
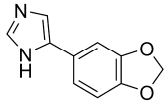
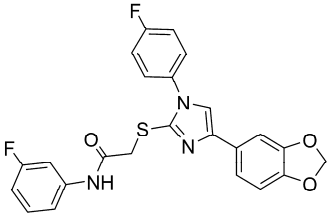
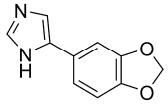
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-53			7756-0340	ChemDiv	57571207
PKL-54			7756-0761	ChemDiv	57571208
PKL-55			7756-0771	ChemDiv	57571209
PKL-56			C177-0121	ChemDiv	57571210
PKL-57			C184-0754	ChemDiv	57571211
PKL-58			C184-0757	ChemDiv	57571212
PKL-59			C184-0761	ChemDiv	57571213

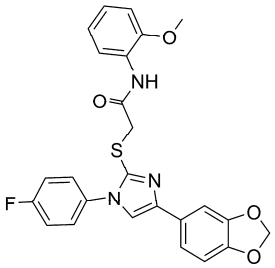
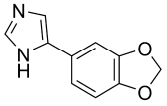
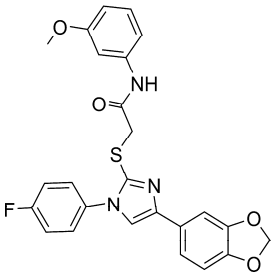
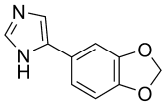
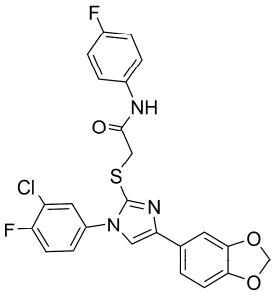
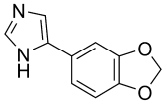
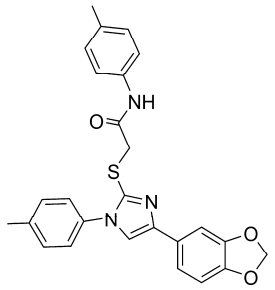
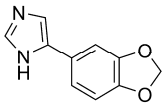
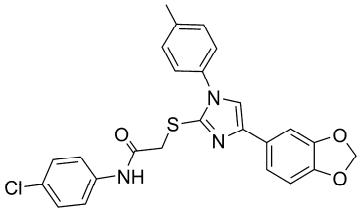
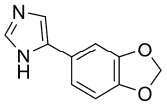
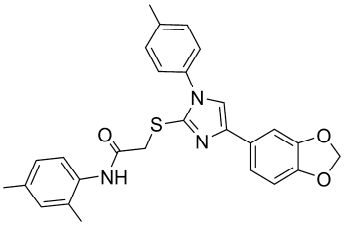
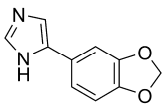
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-60			C184-0762	ChemDiv	57571214
PKL-61			C184-0892	ChemDiv	57571215
PKL-62			C184-0898	ChemDiv	57571216
PKL-63			C184-0918	ChemDiv	57571217
PKL-64			C239-0716	ChemDiv	57571218
PKL-65			D094-0010	ChemDiv	57571142
PKL-66			D094-0011	ChemDiv	57571143

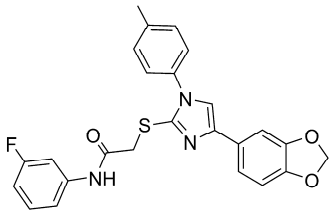
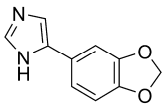
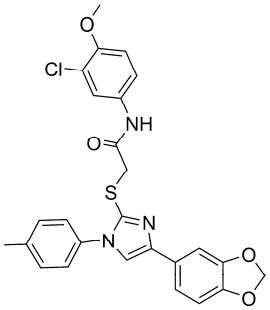
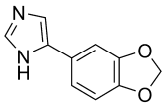
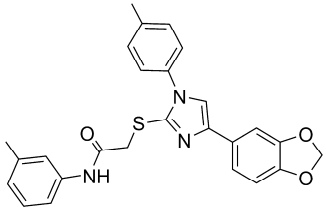
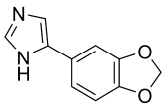
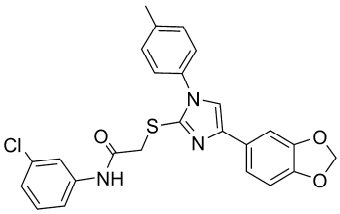
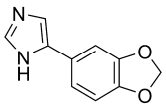
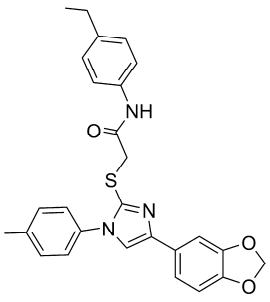
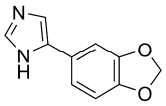
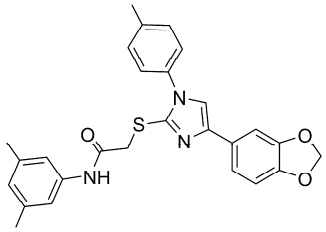
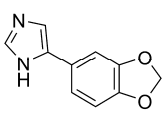
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-67			D094-0013	ChemDiv	57571144
PKL-68			D094-0014	ChemDiv	57571145
PKL-69			D094-0015	ChemDiv	57571146
PKL-70			D094-0016	ChemDiv	57571147
PKL-71			F019-1236	ChemDiv	57571162
PKL-72			F019-1314	ChemDiv	57571163
PKL-73			K242-1013	ChemDiv	57571170
PKL-74			K242-1015	ChemDiv	57571171

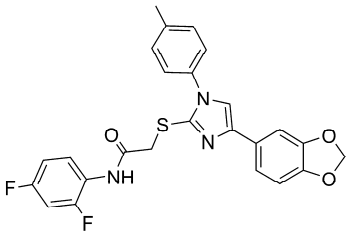
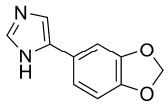
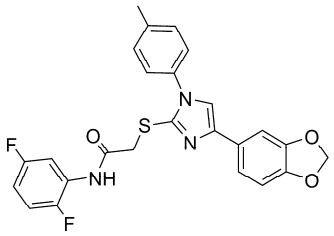
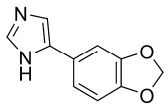
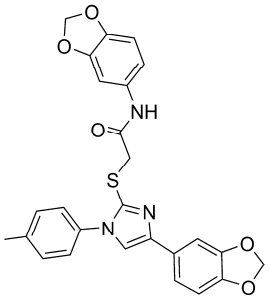
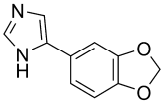
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-75			K242-1019	ChemDiv	57571172
PKL-76			K242-1020	ChemDiv	57571173
PKL-77			K242-1021	ChemDiv	57571174
PKL-78			K242-1037	ChemDiv	57571175
PKL-79			K242-1041	ChemDiv	57571176
PKL-80			K242-1095	ChemDiv	57571177

Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-81			K242-1096	ChemDiv	57571178
PKL-82			K242-1097	ChemDiv	57571179
PKL-83			K242-1098	ChemDiv	57571180
PKL-84			K242-1099	ChemDiv	57571181
PKL-85			K242-1100	ChemDiv	57571182
PKL-86			K242-1101	ChemDiv	57571183

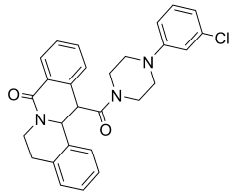
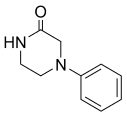
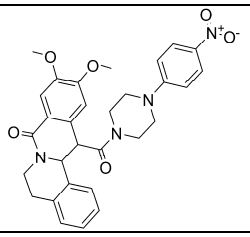
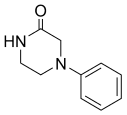
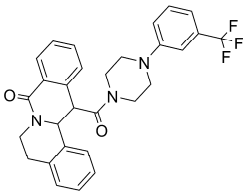
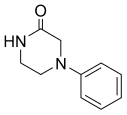
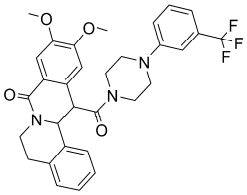
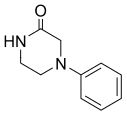
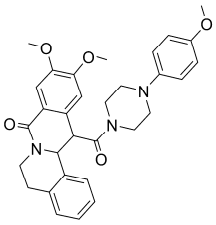
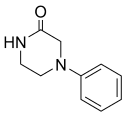
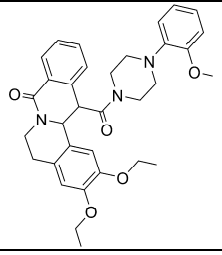
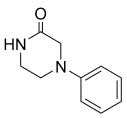
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-87			K242-1102	ChemDiv	57571184
PKL-88			K242-1103	ChemDiv	57571185
PKL-89			K242-1104	ChemDiv	57571186
PKL-90			K242-1108	ChemDiv	57571187
PKL-91			K242-1109	ChemDiv	57571188
PKL-92			K242-1110	ChemDiv	57571189

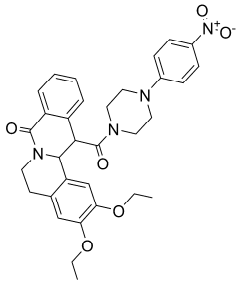
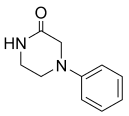
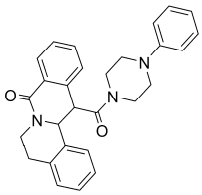
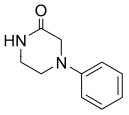
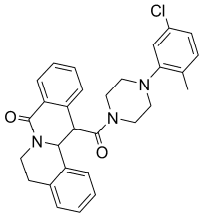
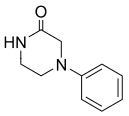
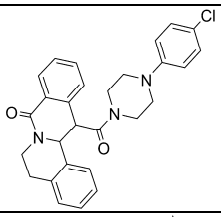
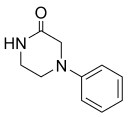
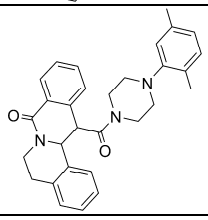
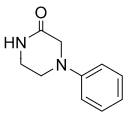
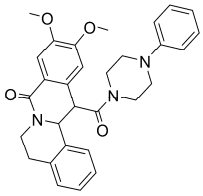
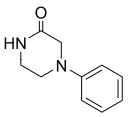
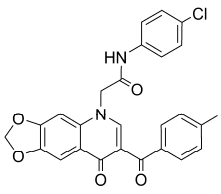
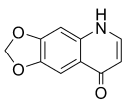
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-93			K242-1111	ChemDiv	57571190
PKL-94			K242-1112	ChemDiv	57571191
PKL-95			K242-1119	ChemDiv	57571192
PKL-96			K242-1126	ChemDiv	57571193
PKL-97			K242-1127	ChemDiv	57571194
PKL-98			K242-1128	ChemDiv	57571195

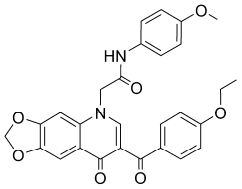
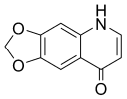
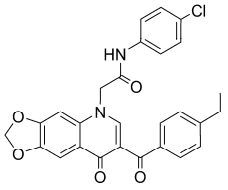
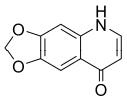
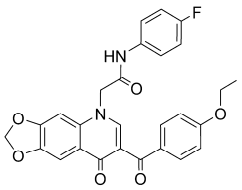
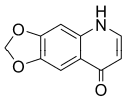
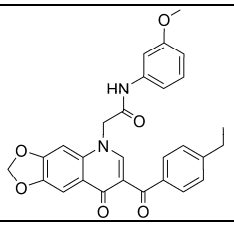
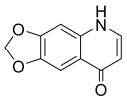
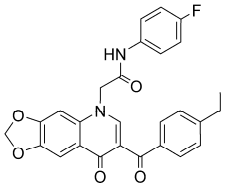
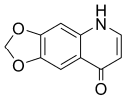
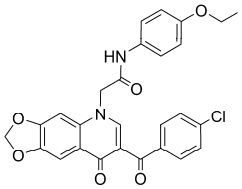
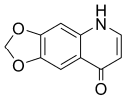
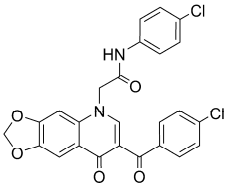
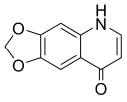
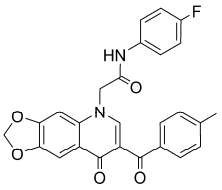
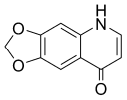
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-99			K242-1129	ChemDiv	57571196
PKL-100			K242-1130	ChemDiv	57571197
PKL-101			K242-1138	ChemDiv	57571198
PKL-102			K242-1139	ChemDiv	57571199
PKL-103			K242-1140	ChemDiv	57571200
PKL-104			K242-1141	ChemDiv	57571201

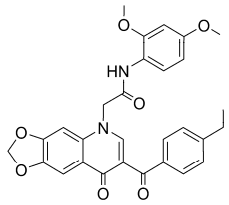
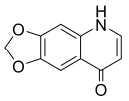
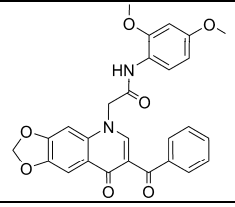
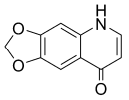
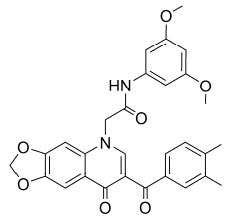
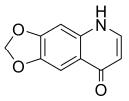
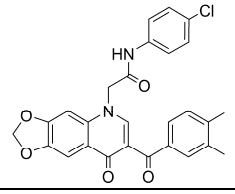
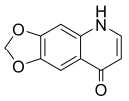
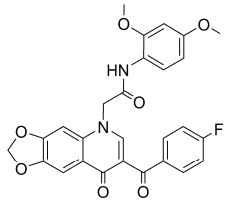
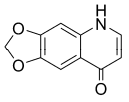
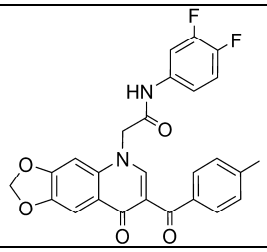
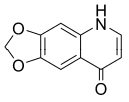
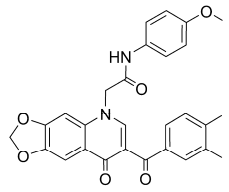
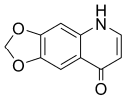
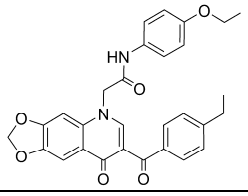
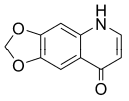
Number	Structure	Virtual scaffold	Vendor cpd. ID	Source	PubChem SID
PKL-105			K242-1142	ChemDiv	57571202
PKL-106			K242-1143	ChemDiv	57571203
PKL-107			K242-1148	ChemDiv	57571204

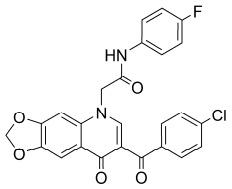
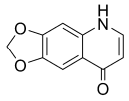
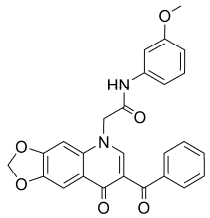
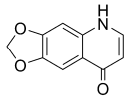
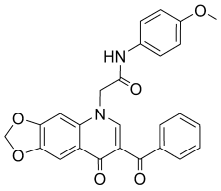
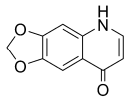
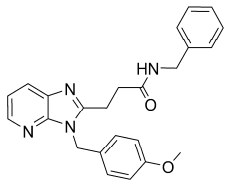
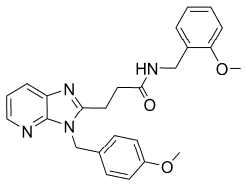
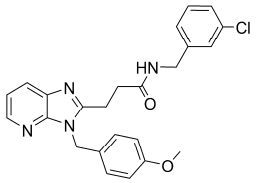
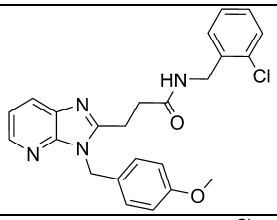
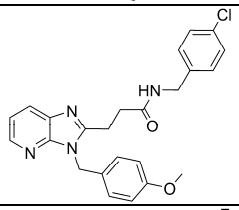
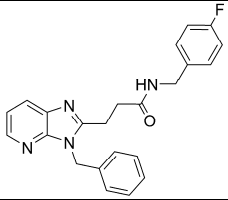
Attachment 2: Table of the 96-membered library of compounds incorporating scaffolds from inactive branches of the pyruvate kinase scaffold tree. The table gives the reference that is used throughout this work, the structures, and the corresponding virtual scaffold as well as the supplier code and the supplier for each compound.

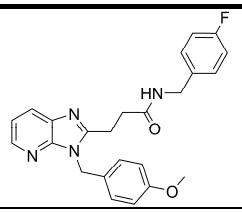
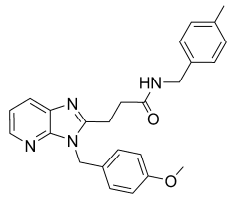
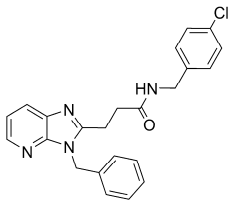
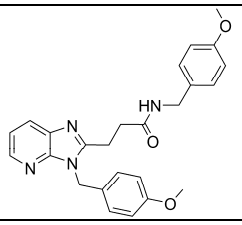
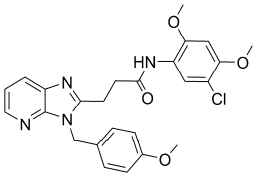
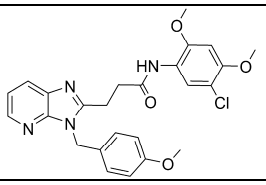
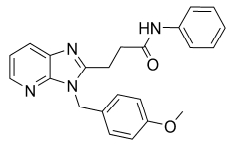
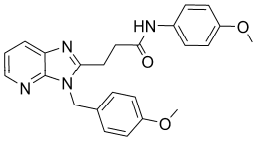
No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-108			C240-0044	ChemDiv
PKL-109			C358-0046	ChemDiv
PKL-110			C240-0169	ChemDiv
PKL-111			C358-0169	ChemDiv
PKL-112			C358-0091	ChemDiv
PKL-113			C289-0037	ChemDiv

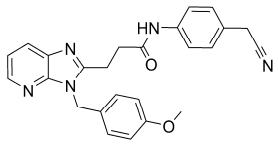
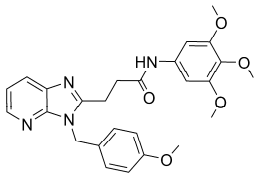
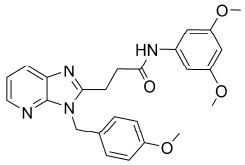
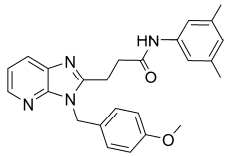
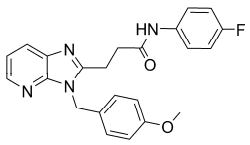
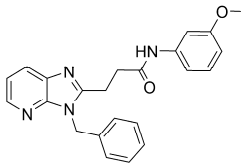
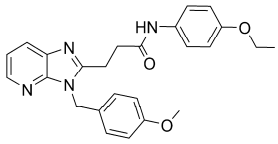
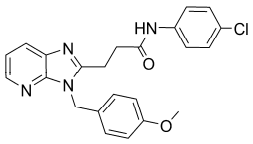
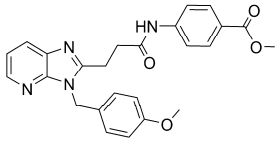
No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-114			C289-0046	ChemDiv
PKL-115			C240-0047	ChemDiv
PKL-116			C240-0170	ChemDiv
PKL-117			C066-2311	ChemDiv
PKL-118			C066-2239	ChemDiv
PKL-119			C066-2160	ChemDiv
PKL-120			C647-0466	ChemDiv

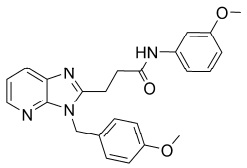
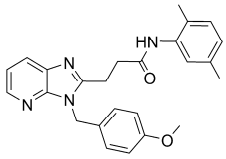
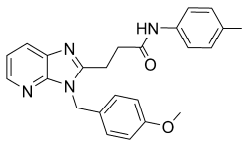
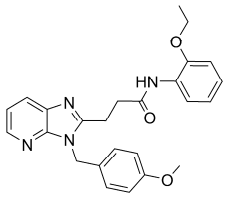
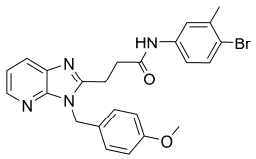
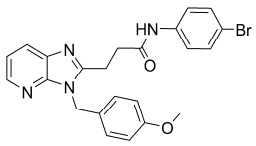
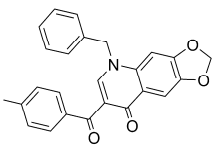
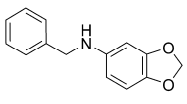
No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-121			C647-0110	ChemDiv
PKL-122			C647-0467	ChemDiv
PKL-123			C647-0390	ChemDiv
PKL-124			C647-0227	ChemDiv
PKL-125			C647-0387	ChemDiv
PKL-126			C647-0192	ChemDiv
PKL-127			C647-0472	ChemDiv
PKL-128			C647-0386	ChemDiv

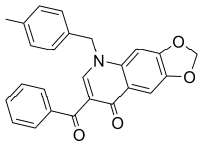
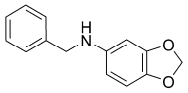
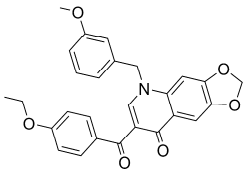
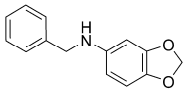
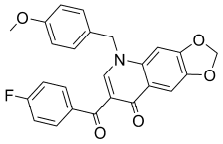
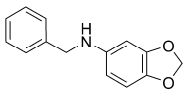
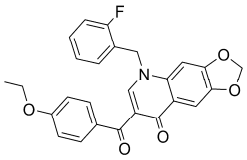
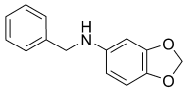
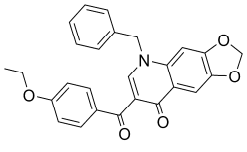
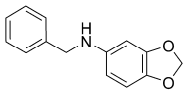
No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-129			C647-0347	ChemDiv
PKL-130			C647-0345	ChemDiv
PKL-131			C647-0268	ChemDiv
PKL-132			C647-0468	ChemDiv
PKL-133			C647-0351	ChemDiv
PKL-134			C647-0426	ChemDiv
PKL-135			C647-0108	ChemDiv
PKL-136			C647-0187	ChemDiv

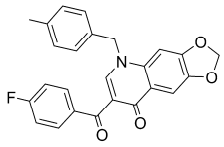
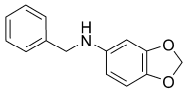
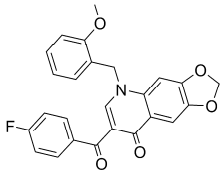
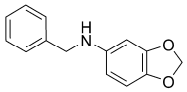
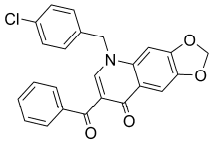
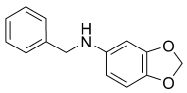
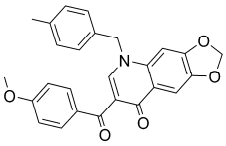
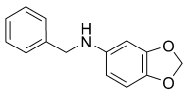
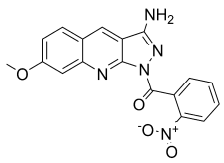
No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-137			C647-0392	ChemDiv
PKL-138			C647-0225	ChemDiv
PKL-139			C647-0105	ChemDiv
PKL-140			C614-0595	ChemDiv
PKL-141			C614-0593	ChemDiv
PKL-142			C614-0921	ChemDiv
PKL-143			C614-0571	ChemDiv
PKL-144			C614-0570	ChemDiv
PKL-145			C614-0985	ChemDiv

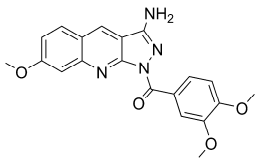
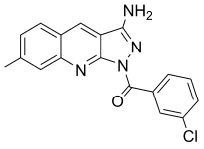
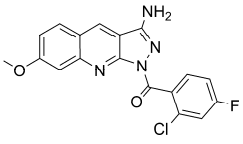
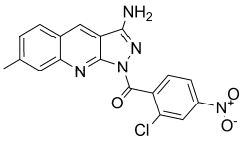
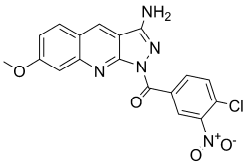
No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-146			C614-0591	ChemDiv
PKL-147			C614-0592	ChemDiv
PKL-148			C614-0979	ChemDiv
PKL-149			C614-0569	ChemDiv
PKL-150			C614-0918	ChemDiv
PKL-151			C614-0923	ChemDiv
PKL-152			C614-0577	ChemDiv
PKL-153			C614-0573	ChemDiv

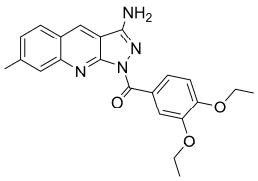
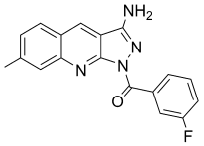
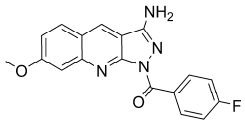
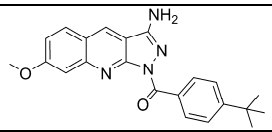
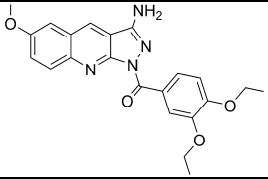
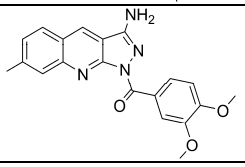
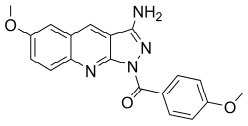
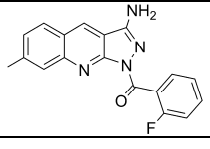
No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-154			C614-0920	ChemDiv
PKL-155			C614-0917	ChemDiv
PKL-156			C614-0913	ChemDiv
PKL-157			C614-0910	ChemDiv
PKL-158			C614-0578	ChemDiv
PKL-159			C614-6013	ChemDiv
PKL-160			C614-0572	ChemDiv
PKL-161			C614-0589	ChemDiv
PKL-162			C614-0933	ChemDiv

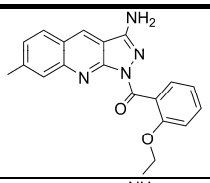
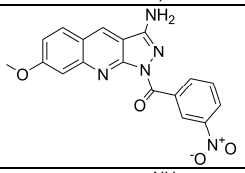
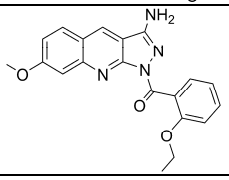
No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-163			C614-0574	ChemDiv
PKL-164			C614-0598	ChemDiv
PKL-165			C614-0907	ChemDiv
PKL-166			C614-0901	ChemDiv
PKL-167			C614-0586	ChemDiv
PKL-168			C614-0935	ChemDiv
PKL-169			C567-0017	ChemDiv

No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-170			C567-0407	ChemDiv
PKL-171			C567-0857	ChemDiv
PKL-172			C567-0967	ChemDiv
PKL-173			C567-0257	ChemDiv
PKL-174			C567-0057	ChemDiv

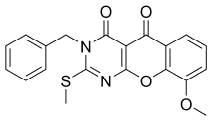
No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-175			C567-0467	ChemDiv
PKL-176			C567-0767	ChemDiv
PKL-177			C567-0107	ChemDiv
PKL-178			C567-0447	ChemDiv
PKL-179			C594-0015	ChemDiv

No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-180			C594-0040	ChemDiv
PKL-181			C619-0133	ChemDiv
PKL-182			C594-0018	ChemDiv
PKL-183			C619-0164	ChemDiv
PKL-184			C594-0036	ChemDiv

No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-185			C619-0208	ChemDiv
PKL-186			C619-0193	ChemDiv
PKL-187			C594-0010	ChemDiv
PKL-188			C066-3140	ChemDiv
PKL-189			C619-0082	ChemDiv
PKL-190			C619-0166	ChemDiv
PKL-191			C660-0117	ChemDiv
PKL-192			C660-0125	ChemDiv

No.	Structure	Virtual Scaffold	Vendor Compound Id	Producer
PKL-193			C619-0202	ChemDiv
PKL-194			C594-0021	ChemDiv
PKL-195			C594-0076	ChemDiv

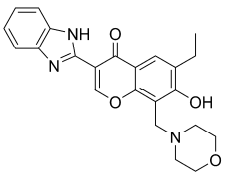
Attachment 3: Table of the 496 members of the γ-pyrone library. The table gives the structures, the reference that is used throughout this work as well as the supplier and the supplier code for each compound. Compounds denoted as “synthesized” have been made by Samy Chammaa-Tortolá and Wolfram Wilk.



GPL-1

InterBioScreen

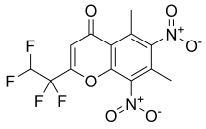
IBS_STOCK6S-12494



GPL-2

InterBioScreen

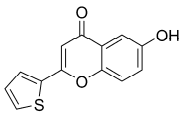
IBS_STOCK1S-01509



GPL-3

InterBioScreen

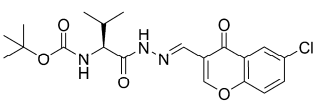
IBS_STOCK2S-49092



GPL-4

InterBioScreen

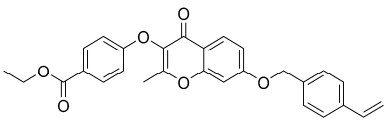
IBS_STOCK3S-37796



GPL-5

InterBioScreen

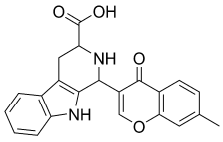
IBS_STOCK4S-88371



GPL-6

InterBioScreen

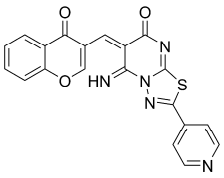
IBS_STOCK1N-42389



GPL-7

InterBioScreen

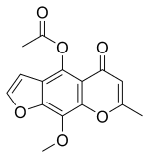
IBS_STOCK1N-24178



GPL-8

InterBioScreen

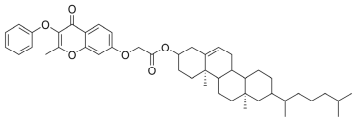
IBS_STOCK5S-59840



GPL-9

InterBioScreen

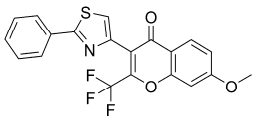
IBS_STOCK1N-32199



GPL-10

InterBioScreen

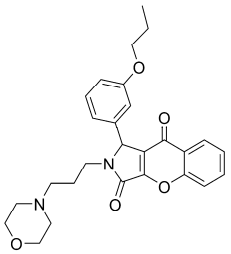
IBS_STOCK1N-65605



GPL-11

InterBioScreen

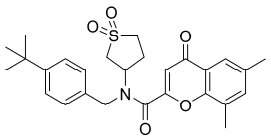
IBS_STOCK1S-21508



GPL-12

InterBioScreen

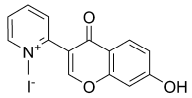
IBS_STOCK4S-44861



GPL-13

InterBioScreen

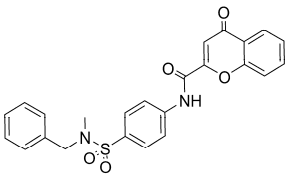
IBS_STOCK5S-26389



GPL-14

InterBioScreen

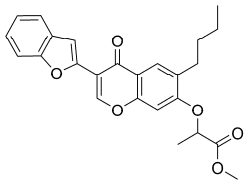
IBS_STOCK1N-41521



GPL-15

InterBioScreen

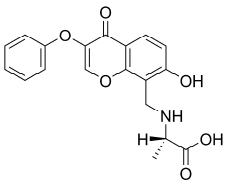
IBS_STOCK5S-88963



GPL-16

InterBioScreen

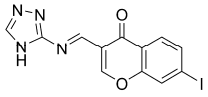
IBS_STOCK5S-29501



GPL-17

InterBioScreen

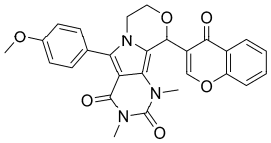
IBS_STOCK1N-01522



GPL-18

InterBioScreen

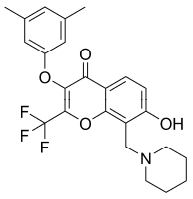
IBS_STOCK3S-15541



GPL-19

InterBioScreen

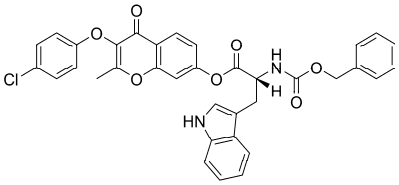
IBS_STOCK5S-89309



GPL-20

InterBioScreen

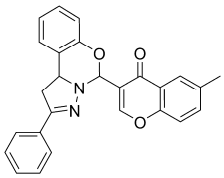
IBS_STOCK3S-25673



GPL-21

InterBioScreen

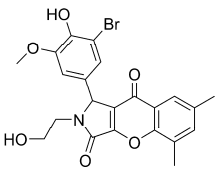
IBS_STOCK1N-11532



GPL-22

InterBioScreen

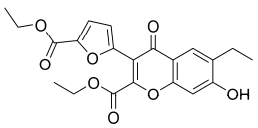
IBS_STOCK2S-27421



GPL-23

InterBioScreen

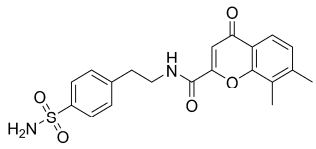
IBS_STOCK5S-55717



GPL-24

InterBioScreen

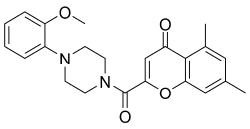
IBS_STOCK5S-37947



GPL-25

InterBioScreen

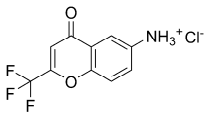
IBS_STOCK5S-95070



GPL-26

InterBioScreen

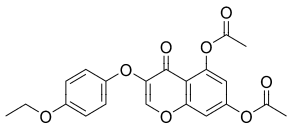
IBS_STOCK5S-92087



GPL-27

InterBioScreen

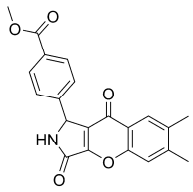
IBS_STOCK2S-37497



GPL-28

InterBioScreen

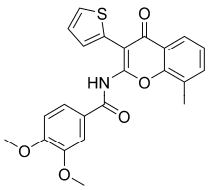
IBS_STOCK1N-00030



GPL-29

InterBioScreen

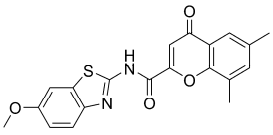
IBS_STOCK5S-46014



GPL-30

InterBioScreen

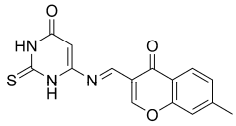
IBS_STOCK6S-03980



GPL-31

InterBioScreen

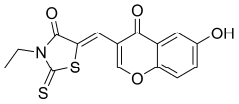
IBS_STOCK5S-51464



GPL-32

InterBioScreen

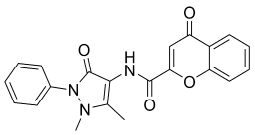
IBS_STOCK3S-72615



GPL-33

InterBioScreen

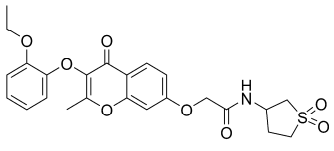
IBS_STOCK5S-28698



GPL-34

InterBioScreen

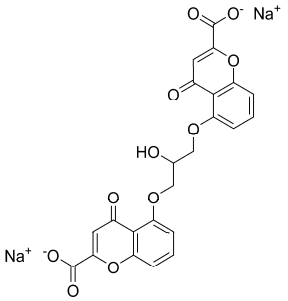
IBS_STOCK5S-93953



GPL-35

InterBioScreen

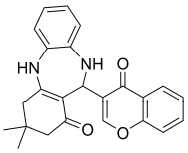
IBS_STOCK6S-40170



GPL-36

InterBioScreen

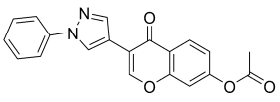
IBS_STOCK1N-03646



GPL-37

InterBioScreen

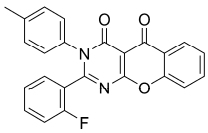
IBS_STOCK3S-93786



GPL-38

InterBioScreen

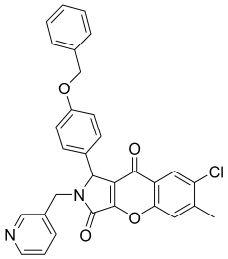
IBS_STOCK3S-00760



GPL-39

InterBioScreen

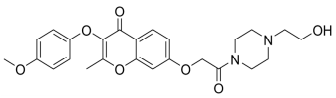
IBS_STOCK5S-27663



GPL-40

InterBioScreen

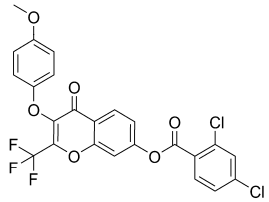
IBS_STOCK5S-33485



GPL-41

InterBioScreen

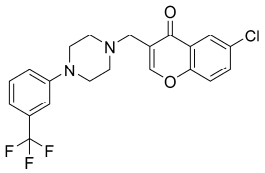
IBS_STOCK6S-38778



GPL-42

InterBioScreen

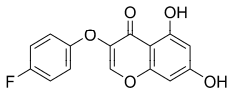
IBS_STOCK2S-61495



GPL-43

InterBioScreen

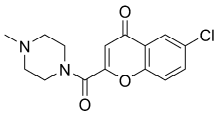
IBS_STOCK1S-94094



GPL-44

InterBioScreen

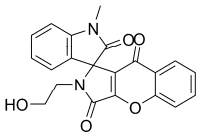
IBS_STOCK1N-00199



GPL-45

InterBioScreen

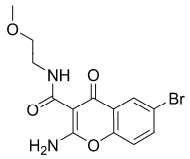
IBS_STOCK5S-84745



GPL-46

InterBioScreen

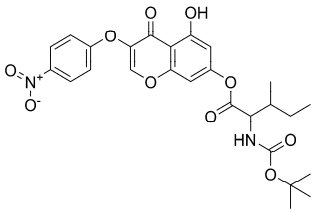
IBS_STOCK5S-35834



GPL-47

InterBioScreen

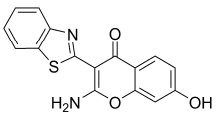
IBS_STOCK6S-19158



GPL-48

InterBioScreen

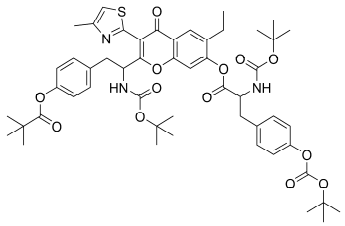
IBS_STOCK1S-58565



GPL-49

InterBioScreen

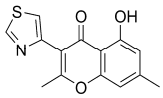
IBS_STOCK1N-28762



GPL-50

InterBioScreen

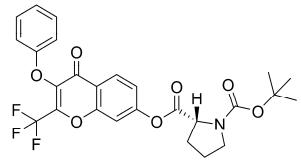
IBS_STOCK1N-01015



GPL-51

InterBioScreen

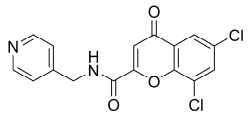
IBS_STOCK1N-06534



GPL-52

InterBioScreen

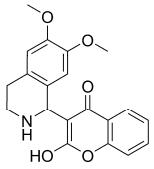
IBS_STOCK1S-62595



GPL-53

InterBioScreen

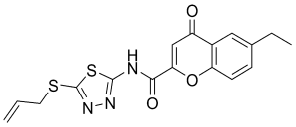
IBS_STOCK5S-85773



GPL-54

InterBioScreen

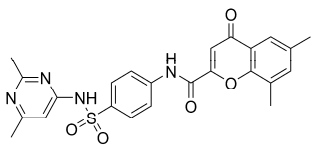
IBS_STOCK1N-68133



GPL-55

InterBioScreen

IBS_STOCK5S-92006



GPL-56

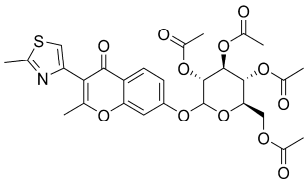
InterBioScreen

IBS_STOCK5S-51355

GPL-57

InterBioScreen

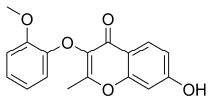
IBS_STOCK1N-25536



GPL-58

InterBioScreen

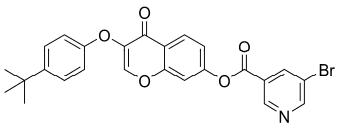
IBS_STOCK1N-09832



GPL-59

InterBioScreen

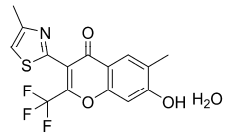
IBS_STOCK1N-38257



GPL-60

InterBioScreen

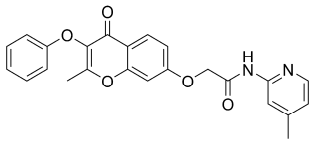
IBS_STOCK2S-71086



GPL-61

InterBioScreen

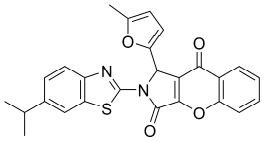
IBS_STOCK1S-64619



GPL-62

InterBioScreen

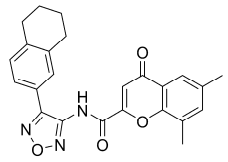
IBS_STOCK6S-42238



GPL-63

InterBioScreen

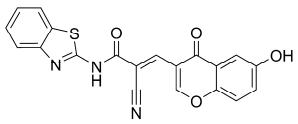
IBS_STOCK5S-83275



GPL-64

InterBioScreen

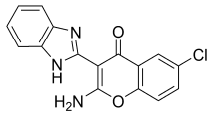
IBS_STOCK6S-30606



GPL-65

InterBioScreen

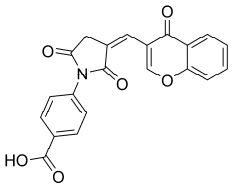
IBS_STOCK5S-34578



GPL-66

InterBioScreen

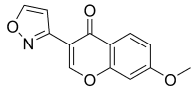
IBS_STOCK6S-17433



GPL-67

InterBioScreen

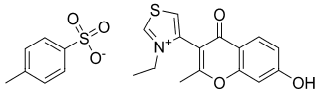
IBS_STOCK6S-40009



GPL-68

InterBioScreen

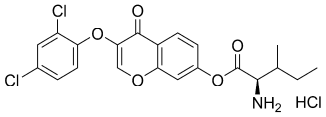
IBS_STOCK1N-05357



GPL-69

InterBioScreen

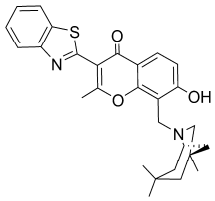
IBS_STOCK5S-54717



GPL-70

InterBioScreen

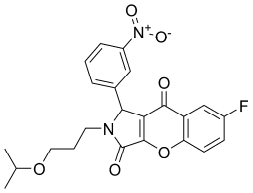
IBS_STOCK1N-44845



GPL-71

InterBioScreen

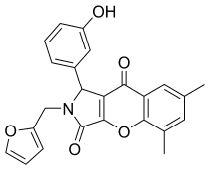
IBS_STOCK1N-29177



GPL-72

InterBioScreen

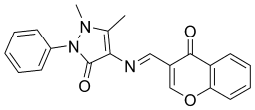
IBS_STOCK4S-40014



GPL-73

InterBioScreen

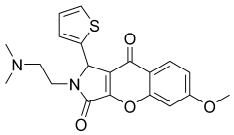
IBS_STOCK5S-35432



GPL-74

InterBioScreen

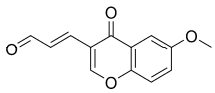
IBS_STOCK2S-85037



GPL-75

InterBioScreen

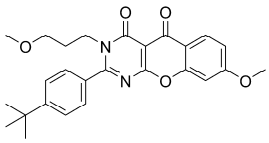
IBS_STOCK5S-90083



GPL-76

InterBioScreen

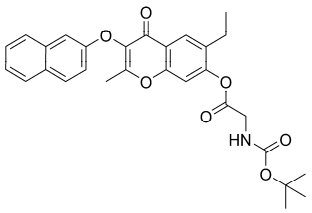
IBS_STOCK5S-39406



GPL-77

InterBioScreen

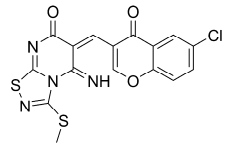
IBS_STOCK5S-28297



GPL-78

InterBioScreen

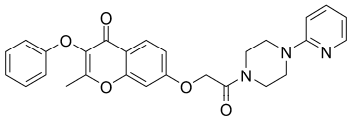
IBS_STOCK1S-08223



GPL-79

InterBioScreen

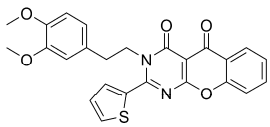
IBS_STOCK5S-47146



GPL-80

InterBioScreen

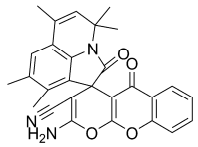
IBS_STOCK6S-35285



GPL-81

InterBioScreen

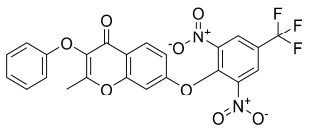
IBS_STOCK5S-41380



GPL-82

InterBioScreen

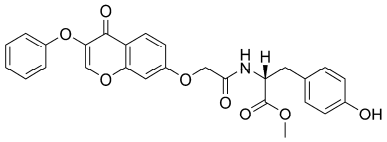
IBS_STOCK5S-40481



GPL-83

InterBioScreen

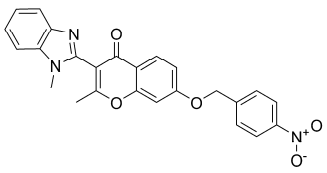
IBS_STOCK1S-58617



GPL-84

InterBioScreen

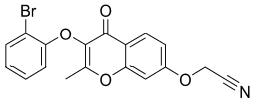
IBS_STOCK1N-08947



GPL-85

InterBioScreen

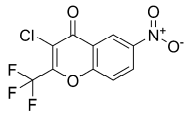
IBS_STOCK1S-67562



GPL-86

InterBioScreen

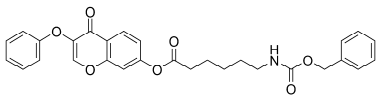
IBS_STOCK4S-64888



GPL-87

InterBioScreen

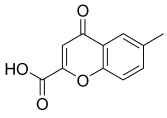
IBS_STOCK2S-44578



GPL-88

InterBioScreen

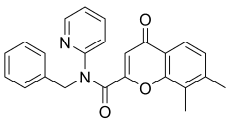
IBS_STOCK5S-45046



GPL-89

InterBioScreen

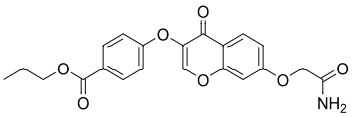
IBS_STOCK1N-15258



GPL-90

InterBioScreen

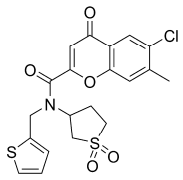
IBS_STOCK5S-89437



GPL-91

InterBioScreen

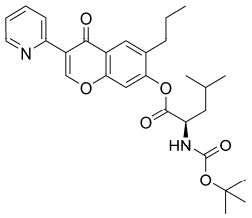
IBS_STOCK4S-71041



GPL-92

InterBioScreen

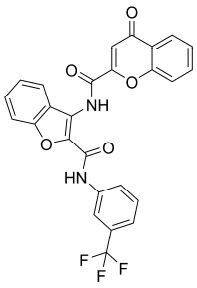
IBS_STOCK5S-47982



GPL-93

InterBioScreen

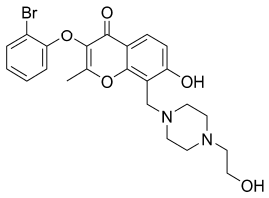
IBS_STOCK1N-42885



GPL-94

InterBioScreen

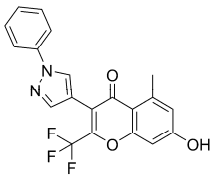
IBS_STOCK5S-89402



GPL-95

InterBioScreen

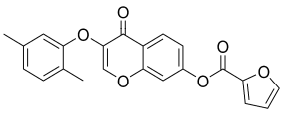
IBS_STOCK4S-79796



GPL-96

InterBioScreen

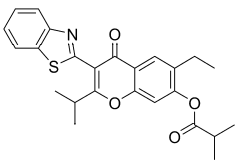
IBS_STOCK1S-17395



GPL-97

InterBioScreen

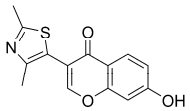
IBS_STOCK1N-30832



GPL-98

InterBioScreen

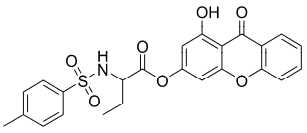
IBS_STOCK1N-29077



GPL-99

InterBioScreen

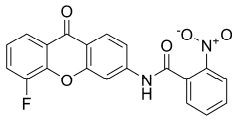
IBS_STOCK1N-01163



GPL-100

InterBioScreen

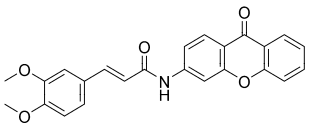
IBS_STOCK6S-00457



GPL-101

InterBioScreen

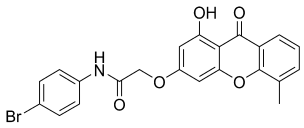
IBS_STOCK5S-92398



GPL-102

InterBioScreen

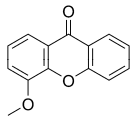
IBS_STOCK5S-92164



GPL-103

InterBioScreen

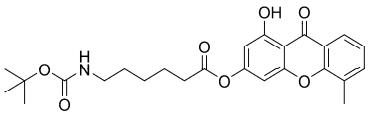
IBS_STOCK6S-00695



GPL-104

InterBioScreen

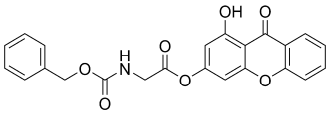
IBS_STOCK1N-29323



GPL-105

InterBioScreen

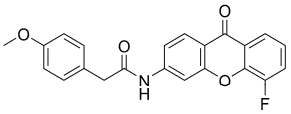
IBS_STOCK6S-01602



GPL-106

InterBioScreen

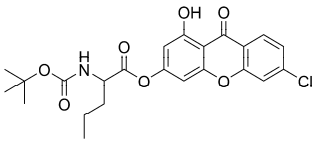
IBS_STOCK6S-00865



GPL-107

InterBioScreen

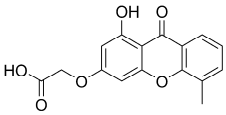
IBS_STOCK5S-96215



GPL-108

InterBioScreen

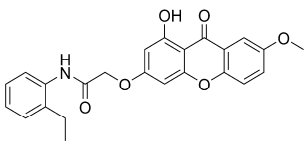
IBS_STOCK5S-99946



GPL-109

InterBioScreen

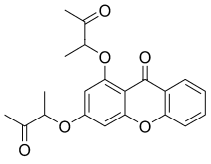
IBS_STOCK1N-64226



GPL-110

InterBioScreen

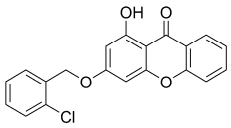
IBS_STOCK6S-36046



GPL-111

InterBioScreen

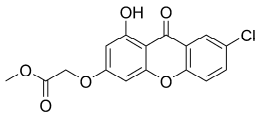
IBS_STOCK1N-58373



GPL-112

InterBioScreen

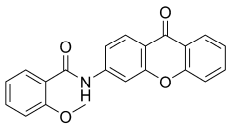
IBS_STOCK5S-56038



GPL-113

InterBioScreen

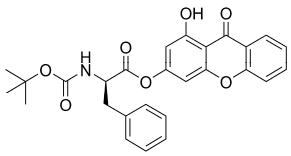
IBS_STOCK5S-90008



GPL-114

InterBioScreen

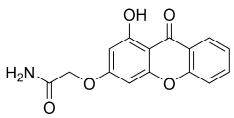
IBS_STOCK5S-84939



GPL-115

InterBioScreen

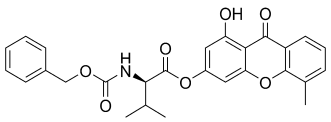
IBS_STOCK6S-01706



GPL-116

InterBioScreen

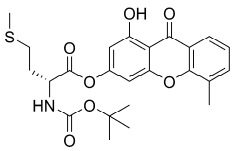
IBS_STOCK1N-65252



GPL-117

InterBioScreen

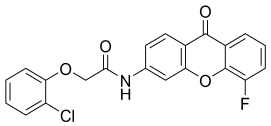
IBS_STOCK6S-01271



GPL-118

InterBioScreen

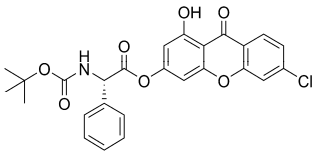
IBS_STOCK5S-98367



GPL-119

InterBioScreen

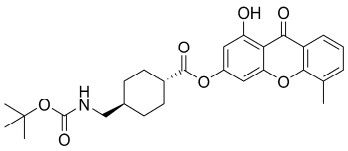
IBS_STOCK5S-96225



GPL-120

InterBioScreen

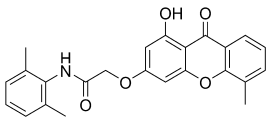
IBS_STOCK5S-96600



GPL-121

InterBioScreen

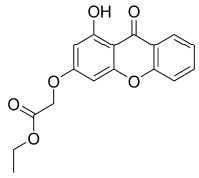
IBS_STOCK5S-99381



GPL-122

InterBioScreen

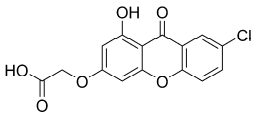
IBS_STOCK5S-98469



GPL-123

InterBioScreen

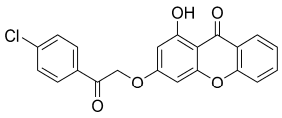
IBS_STOCK1N-54982



GPL-124

InterBioScreen

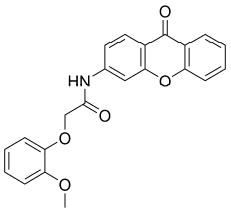
IBS_STOCK5S-84903



GPL-125

InterBioScreen

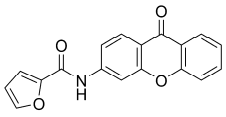
IBS_STOCK1N-58144



GPL-126

InterBioScreen

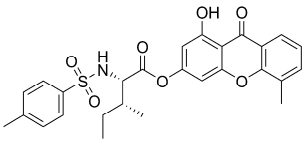
IBS_STOCK5S-83626



GPL-127

InterBioScreen

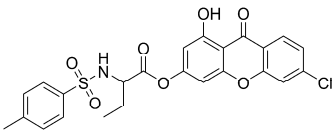
IBS_STOCK5S-83537



GPL-128

InterBioScreen

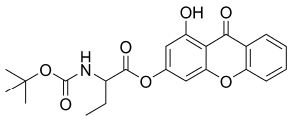
IBS_STOCK6S-01209



GPL-129

InterBioScreen

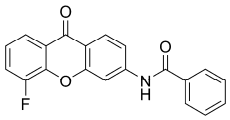
IBS_STOCK5S-99796



GPL-130

InterBioScreen

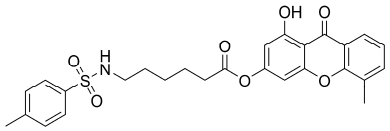
IBS_STOCK6S-01205



GPL-131

InterBioScreen

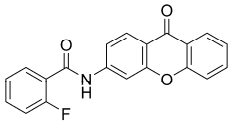
IBS_STOCK5S-85251



GPL-132

InterBioScreen

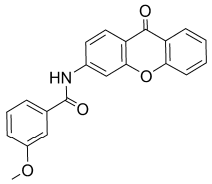
IBS_STOCK5S-96967



GPL-133

InterBioScreen

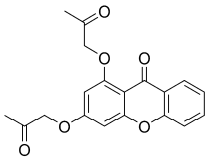
IBS_STOCK5S-87546



GPL-134

InterBioScreen

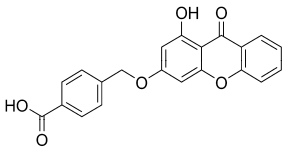
IBS_STOCK5S-92982



GPL-135

InterBioScreen

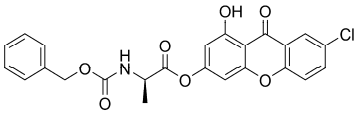
IBS_STOCK1N-56315



GPL-136

InterBioScreen

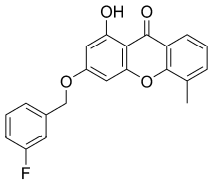
IBS_STOCK1N-57878



GPL-137

InterBioScreen

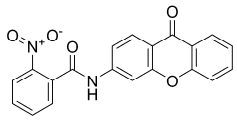
IBS_STOCK5S-98312



GPL-138

InterBioScreen

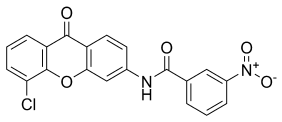
IBS_STOCK5S-73490



GPL-139

InterBioScreen

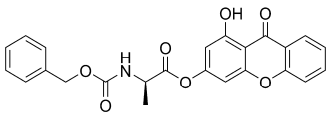
IBS_STOCK5S-84641



GPL-140

InterBioScreen

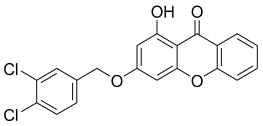
IBS_STOCK5S-86097



GPL-141

InterBioScreen

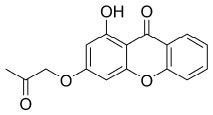
IBS_STOCK5S-99107



GPL-142

InterBioScreen

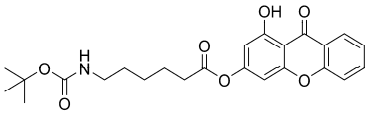
IBS_STOCK5S-26258



GPL-143

InterBioScreen

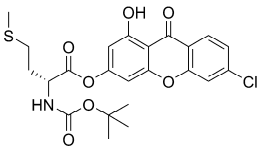
IBS_STOCK1N-64473



GPL-144

InterBioScreen

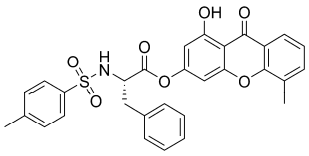
IBS_STOCK5S-98801



GPL-145

InterBioScreen

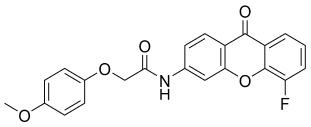
IBS_STOCK6S-01054



GPL-146

InterBioScreen

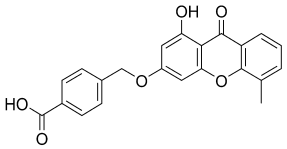
IBS_STOCK6S-01043



GPL-147

InterBioScreen

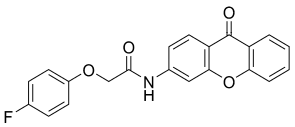
IBS_STOCK5S-95288



GPL-148

InterBioScreen

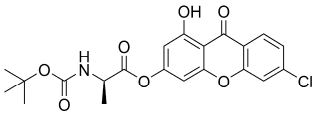
IBS_STOCK1N-59040



GPL-149

InterBioScreen

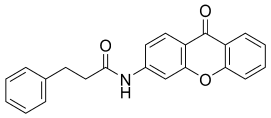
IBS_STOCK5S-90338



GPL-150

InterBioScreen

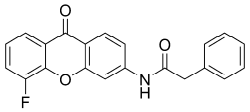
IBS_STOCK5S-98502



GPL-151

InterBioScreen

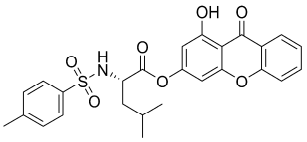
IBS_STOCK5S-95617



GPL-152

InterBioScreen

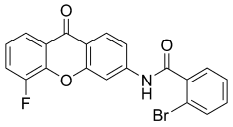
IBS_STOCK5S-85596



GPL-153

InterBioScreen

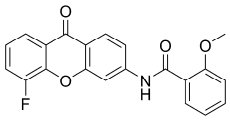
IBS_STOCK6S-00191



GPL-154

InterBioScreen

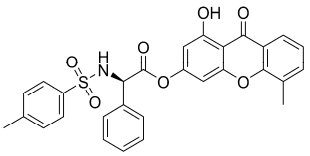
IBS_STOCK5S-95478



GPL-155

InterBioScreen

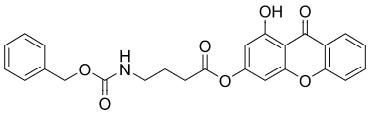
IBS_STOCK5S-91397



GPL-156

InterBioScreen

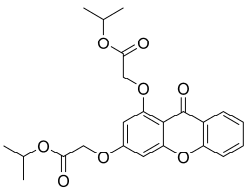
IBS_STOCK5S-98508



GPL-157

InterBioScreen

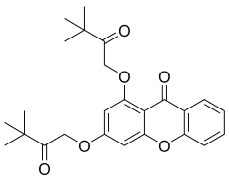
IBS_STOCK5S-99090



GPL-158

InterBioScreen

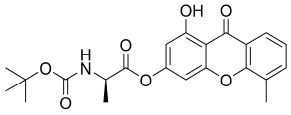
IBS_STOCK1N-57111



GPL-159

InterBioScreen

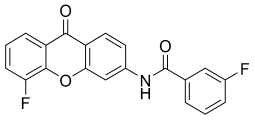
IBS_STOCK5S-32922



GPL-160

InterBioScreen

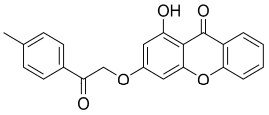
IBS_STOCK5S-99027



GPL-161

InterBioScreen

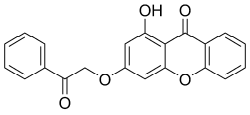
IBS_STOCK5S-87045



GPL-162

InterBioScreen

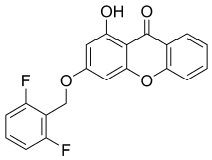
IBS_STOCK1N-56619



GPL-163

InterBioScreen

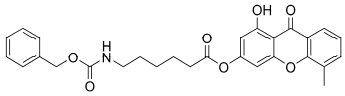
IBS_STOCK1N-56203



GPL-164

InterBioScreen

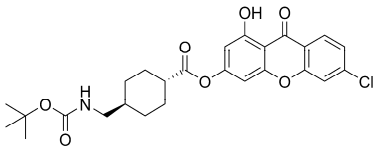
IBS_STOCK5S-29102



GPL-165

InterBioScreen

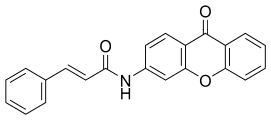
IBS_STOCK5S-96848



GPL-166

InterBioScreen

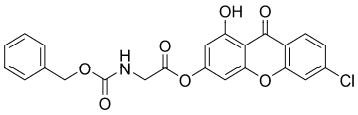
IBS_STOCK5S-97461



GPL-167

InterBioScreen

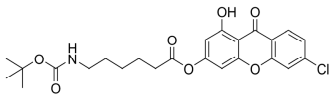
IBS_STOCK5S-90886



GPL-168

InterBioScreen

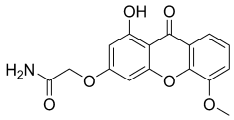
IBS_STOCK6S-00399



GPL-169

InterBioScreen

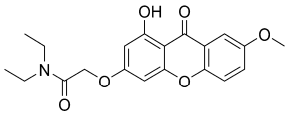
IBS_STOCK5S-96688



GPL-170

InterBioScreen

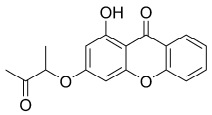
IBS_STOCK1N-64644



GPL-171

InterBioScreen

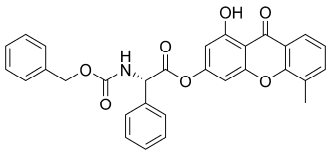
IBS_STOCK1N-69290



GPL-172

InterBioScreen

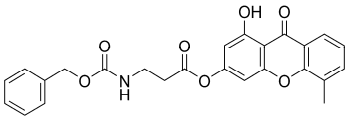
IBS_STOCK1N-64332



GPL-173

InterBioScreen

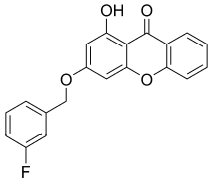
IBS_STOCK6S-00937



GPL-174

InterBioScreen

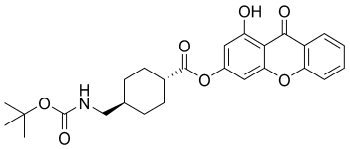
IBS_STOCK1N-60281



GPL-175

InterBioScreen

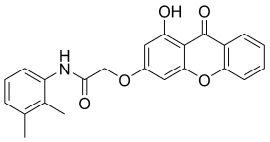
IBS_STOCK5S-38709



GPL-176

InterBioScreen

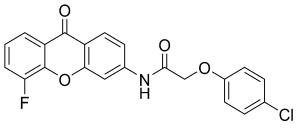
IBS_STOCK5S-99506



GPL-177

InterBioScreen

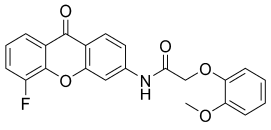
IBS_STOCK6S-32273



GPL-178

InterBioScreen

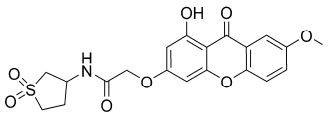
IBS_STOCK5S-90043



GPL-179

InterBioScreen

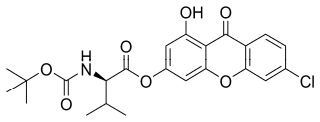
IBS_STOCK5S-84976



GPL-180

InterBioScreen

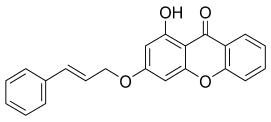
IBS_STOCK6S-38432



GPL-181

InterBioScreen

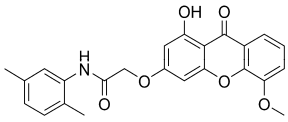
IBS_STOCK5S-98286



GPL-182

InterBioScreen

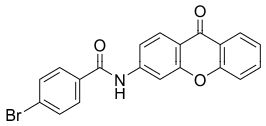
IBS_STOCK1N-58590



GPL-183

InterBioScreen

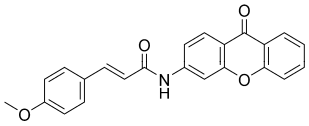
IBS_STOCK6S-33822



GPL-184

InterBioScreen

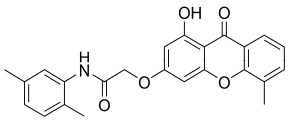
IBS_STOCK5S-92000



GPL-185

InterBioScreen

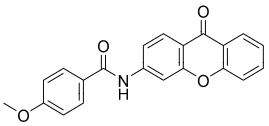
IBS_STOCK5S-85945



GPL-186

InterBioScreen

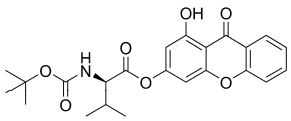
IBS_STOCK5S-96652



GPL-187

InterBioScreen

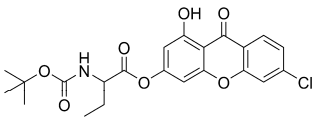
IBS_STOCK5S-89989



GPL-188

InterBioScreen

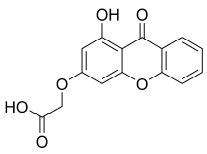
IBS_STOCK6S-00046



GPL-189

InterBioScreen

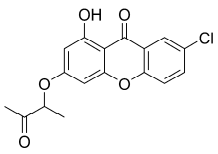
IBS_STOCK5S-99038



GPL-190

InterBioScreen

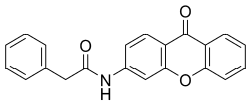
IBS_STOCK1N-57056



GPL-191

InterBioScreen

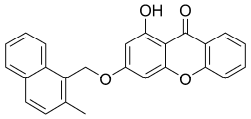
IBS_STOCK5S-87118



GPL-192

InterBioScreen

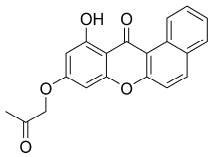
IBS_STOCK5S-84472



GPL-193

InterBioScreen

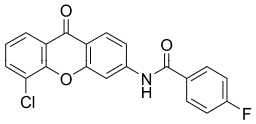
IBS_STOCK5S-46772



GPL-194

InterBioScreen

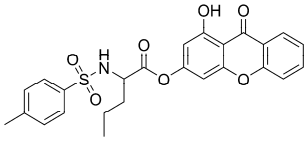
IBS_STOCK5S-93582



GPL-195

InterBioScreen

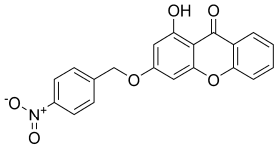
IBS_STOCK5S-89992



GPL-196

InterBioScreen

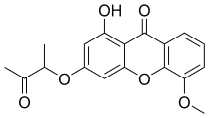
IBS_STOCK6S-01557



GPL-197

InterBioScreen

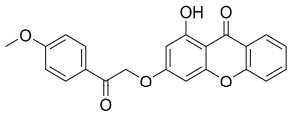
IBS_STOCK5S-56240



GPL-198

InterBioScreen

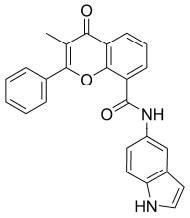
IBS_STOCK1N-64331



GPL-199

InterBioScreen

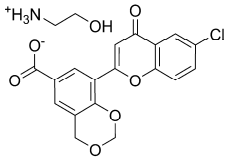
IBS_STOCK1N-57983



GPL-200

InterBioScreen

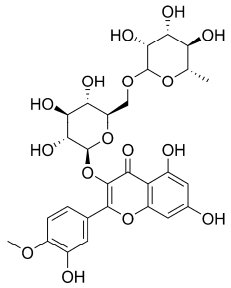
IBS_STOCK1N-69379



GPL-201

InterBioScreen

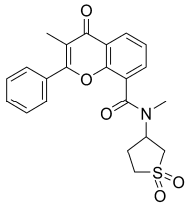
IBS_STOCK1N-23953



GPL-202

InterBioScreen

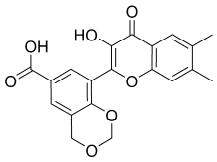
IBS_STOCK1N-54280



GPL-203

InterBioScreen

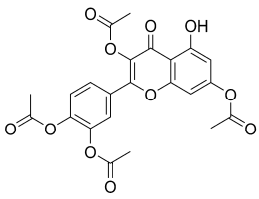
IBS_STOCK6S-22547



GPL-204

InterBioScreen

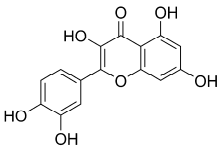
IBS_STOCK1N-02575



GPL-205

InterBioScreen

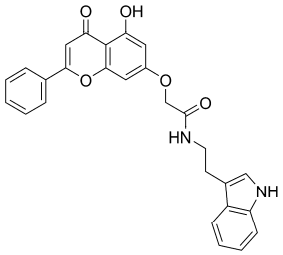
IBS_STOCK1N-00292



GPL-206

InterBioScreen

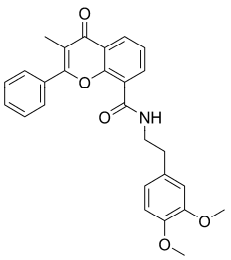
IBS_STOCK1N-04222



GPL-207

InterBioScreen

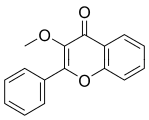
IBS_STOCK1N-69016



GPL-208

InterBioScreen

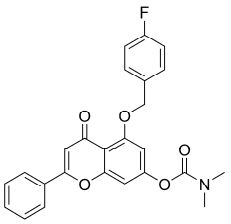
IBS_STOCK6S-19875



GPL-209

InterBioScreen

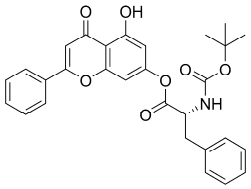
IBS_STOCK1N-24546



GPL-210

InterBioScreen

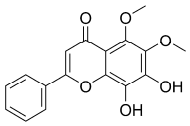
IBS_STOCK6S-27528



GPL-211

InterBioScreen

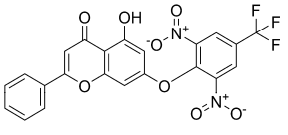
IBS_STOCK1N-11907



GPL-212

InterBioScreen

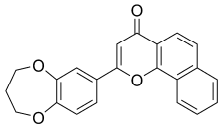
IBS_STOCK1N-68204



GPL-213

InterBioScreen

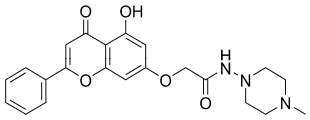
IBS_STOCK1S-49140



GPL-214

InterBioScreen

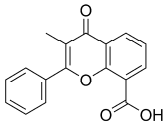
IBS_STOCK5S-58225



GPL-215

InterBioScreen

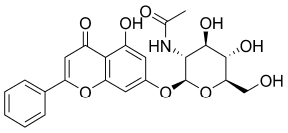
IBS_STOCK6S-42828



GPL-216

InterBioScreen

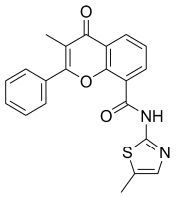
IBS_STOCK1N-69951



GPL-217

InterBioScreen

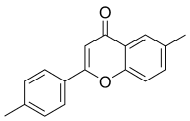
IBS_STOCK1N-59846



GPL-218

InterBioScreen

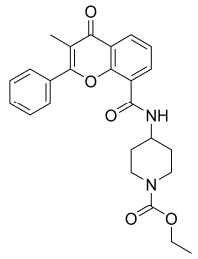
IBS_STOCK6S-26993



GPL-219

InterBioScreen

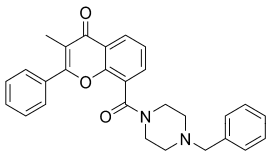
IBS_STOCK1N-45047



GPL-220

InterBioScreen

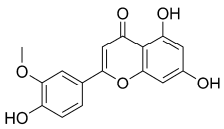
IBS_STOCK6S-22000



GPL-221

InterBioScreen

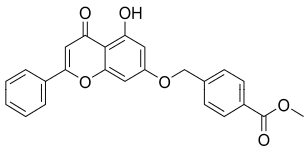
IBS_STOCK6S-25437



GPL-222

InterBioScreen

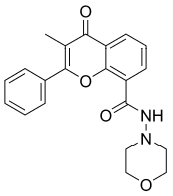
IBS_STOCK1N-14981



GPL-223

InterBioScreen

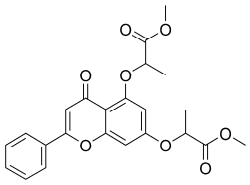
IBS_STOCK1N-22252



GPL-224

InterBioScreen

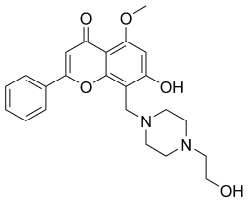
IBS_STOCK6S-23803



GPL-225

InterBioScreen

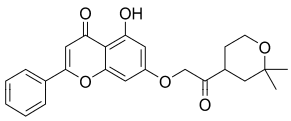
IBS_STOCK1N-12184



GPL-226

InterBioScreen

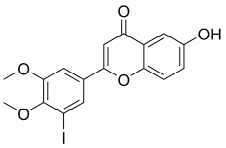
IBS_STOCK6S-27548



GPL-227

InterBioScreen

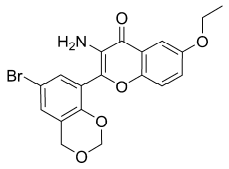
IBS_STOCK1N-01253



GPL-228

InterBioScreen

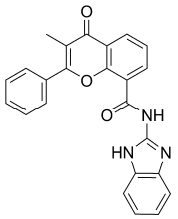
IBS_STOCK1N-41708



GPL-229

InterBioScreen

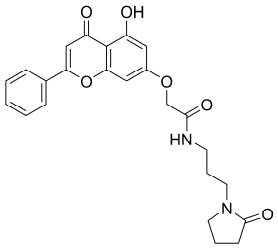
IBS_STOCK1N-08722



GPL-230

InterBioScreen

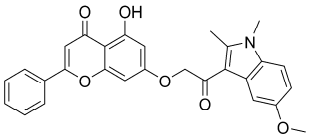
IBS_STOCK6S-21081



GPL-231

InterBioScreen

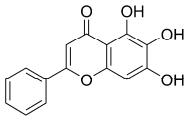
IBS_STOCK1N-69353



GPL-232

InterBioScreen

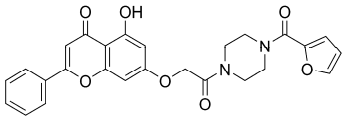
IBS_STOCK1N-50710



GPL-233

InterBioScreen

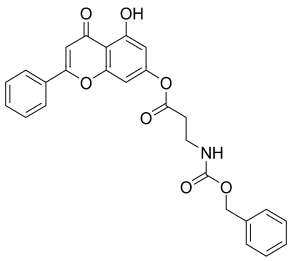
IBS_STOCK1N-28559



GPL-234

InterBioScreen

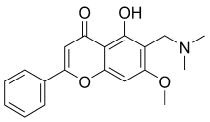
IBS_STOCK6S-40325



GPL-235

InterBioScreen

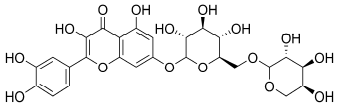
IBS_STOCK1N-45295



GPL-236

InterBioScreen

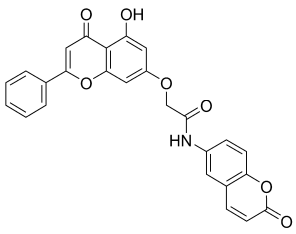
IBS_STOCK1N-69484



GPL-237

InterBioScreen

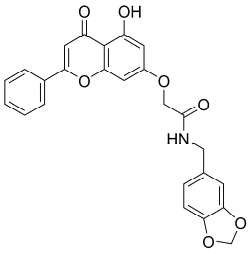
IBS_STOCK1N-50334



GPL-238

InterBioScreen

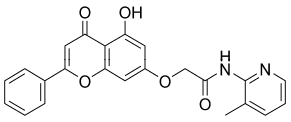
IBS_STOCK1N-69858



GPL-239

InterBioScreen

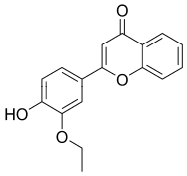
IBS_STOCK6S-43461



GPL-240

InterBioScreen

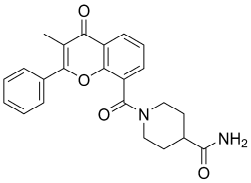
IBS_STOCK6S-40252



GPL-241

InterBioScreen

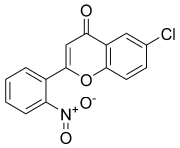
IBS_STOCK1N-33002



GPL-242

InterBioScreen

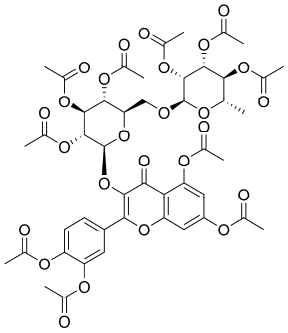
IBS_STOCK6S-25683



GPL-243

InterBioScreen

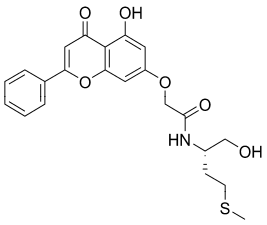
IBS_STOCK1S-80476



GPL-244

InterBioScreen

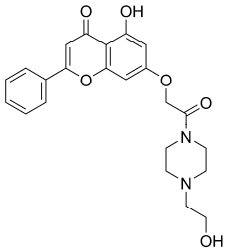
IBS_STOCK1N-17598



GPL-245

InterBioScreen

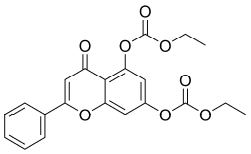
IBS_STOCK1N-70623



GPL-246

InterBioScreen

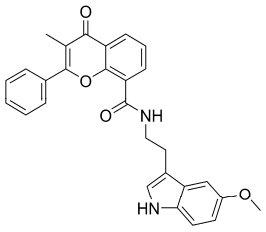
IBS_STOCK1N-69552



GPL-247

InterBioScreen

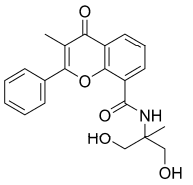
IBS_STOCK1N-45425



GPL-248

InterBioScreen

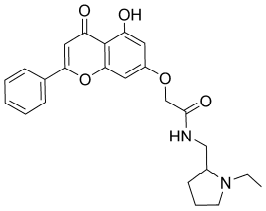
IBS_STOCK1N-69231



GPL-249

InterBioScreen

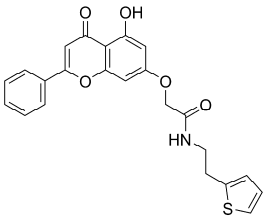
IBS_STOCK1N-70113



GPL-250

InterBioScreen

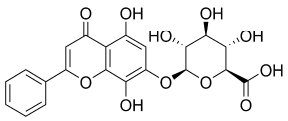
IBS_STOCK1N-68637



GPL-251

InterBioScreen

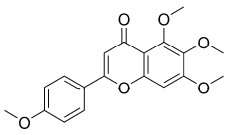
IBS_STOCK6S-37597



GPL-252

InterBioScreen

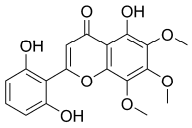
IBS_STOCK1N-52965



GPL-253

InterBioScreen

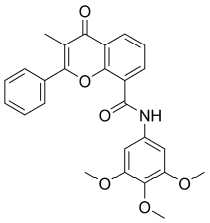
IBS_STOCK1N-09778



GPL-254

InterBioScreen

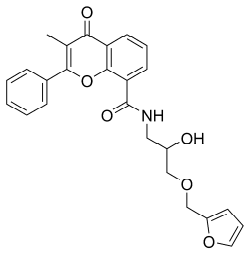
IBS_STOCK1N-17641



GPL-255

InterBioScreen

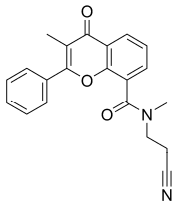
IBS_STOCK6S-21753



GPL-256

InterBioScreen

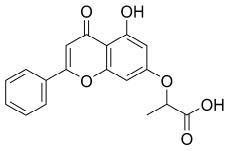
IBS_STOCK1N-70036



GPL-257

InterBioScreen

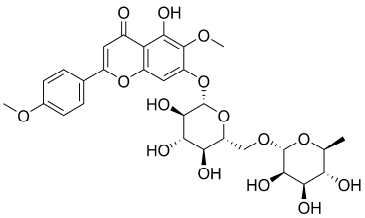
IBS_STOCK6S-21383



GPL-258

InterBioScreen

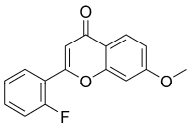
IBS_STOCK1N-09647



GPL-259

InterBioScreen

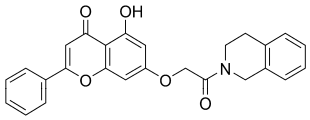
IBS_STOCK1N-08706



GPL-260

InterBioScreen

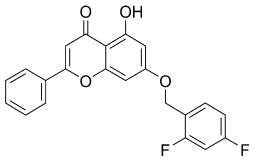
IBS_STOCK3S-76041



GPL-261

InterBioScreen

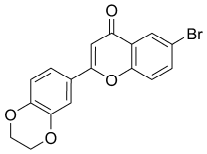
IBS_STOCK6S-36093



GPL-262

InterBioScreen

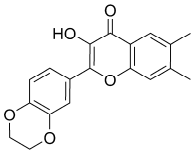
IBS_STOCK4S-89839



GPL-263

InterBioScreen

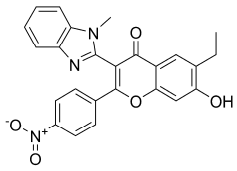
IBS_STOCK1N-11094



GPL-264

InterBioScreen

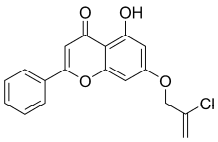
IBS_STOCK1N-08520



GPL-265

InterBioScreen

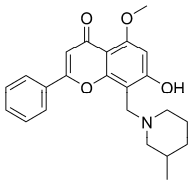
IBS_STOCK1S-75174



GPL-266

InterBioScreen

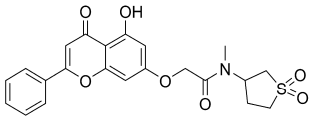
IBS_STOCK1N-47594



GPL-267

InterBioScreen

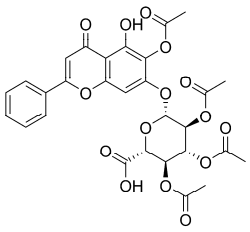
IBS_STOCK6S-20742



GPL-268

InterBioScreen

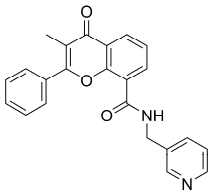
IBS_STOCK6S-16843



GPL-269

InterBioScreen

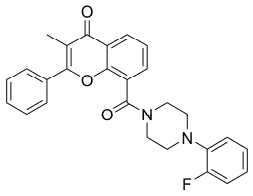
IBS_STOCK1N-03986



GPL-270

InterBioScreen

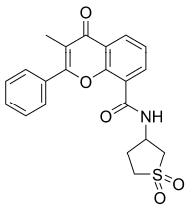
IBS_STOCK6S-24693



GPL-271

InterBioScreen

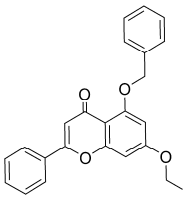
IBS_STOCK6S-22721



GPL-272

InterBioScreen

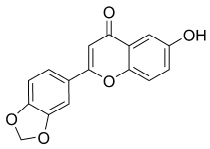
IBS_STOCK6S-19565



GPL-273

InterBioScreen

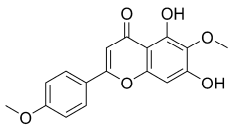
IBS_STOCK5S-50237



GPL-274

InterBioScreen

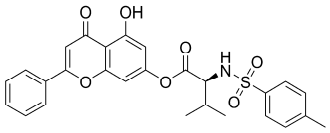
IBS_STOCK1N-33354



GPL-275

InterBioScreen

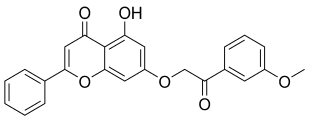
IBS_STOCK1N-09903



GPL-276

InterBioScreen

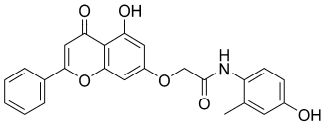
IBS_STOCK1N-10107



GPL-277

InterBioScreen

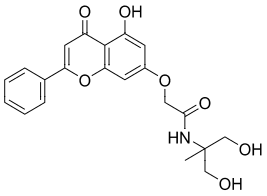
IBS_STOCK1N-54004



GPL-278

InterBioScreen

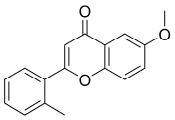
IBS_STOCK1N-68651



GPL-279

InterBioScreen

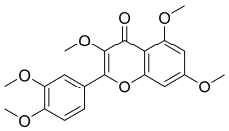
IBS_STOCK6S-40664



GPL-280

InterBioScreen

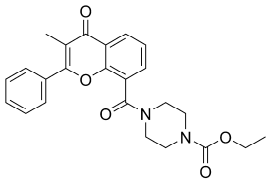
IBS_STOCK5S-29346



GPL-281

InterBioScreen

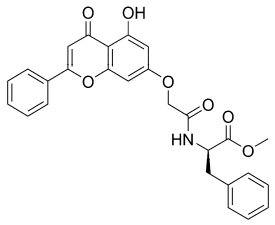
IBS_STOCK1N-00418



GPL-282

InterBioScreen

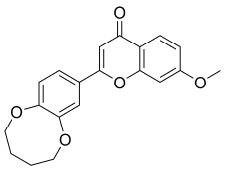
IBS_STOCK6S-32276



GPL-283

InterBioScreen

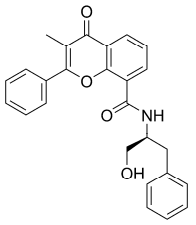
IBS_STOCK1N-69570



GPL-284

InterBioScreen

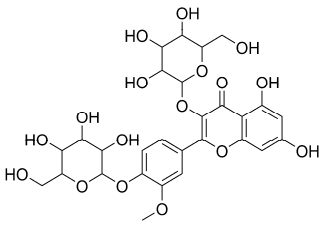
IBS_STOCK1N-00166



GPL-285

InterBioScreen

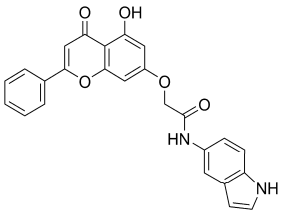
IBS_STOCK1N-69893



GPL-286

InterBioScreen

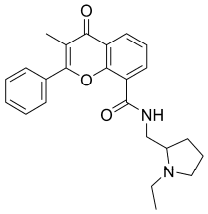
IBS_STOCK1N-08230



GPL-287

InterBioScreen

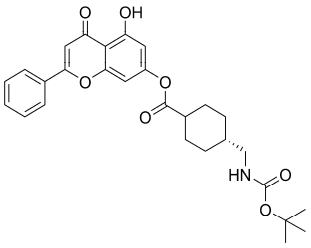
IBS_STOCK1N-70703



GPL-288

InterBioScreen

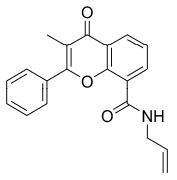
IBS_STOCK6S-25601



GPL-289

InterBioScreen

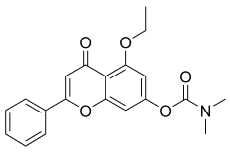
IBS_STOCK1N-54487



GPL-290

InterBioScreen

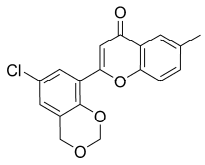
IBS_STOCK6S-28611



GPL-291

InterBioScreen

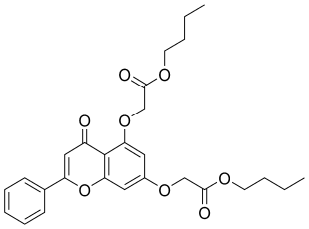
IBS_STOCK6S-29564



GPL-292

InterBioScreen

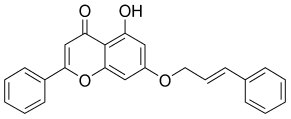
IBS_STOCK1N-00284



GPL-293

InterBioScreen

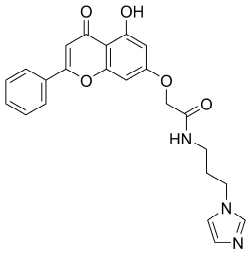
IBS_STOCK1N-07623



GPL-294

InterBioScreen

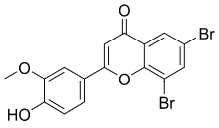
IBS_STOCK1N-22287



GPL-295

InterBioScreen

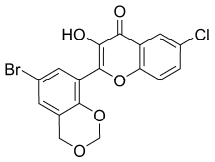
IBS_STOCK1N-69175



GPL-296

InterBioScreen

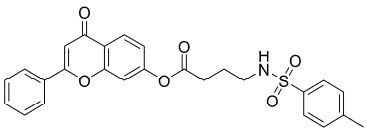
IBS_STOCK4S-95609



GPL-297

InterBioScreen

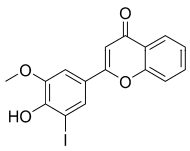
IBS_STOCK1N-07662



GPL-298

InterBioScreen

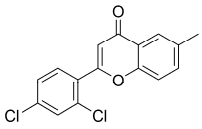
IBS_STOCK1N-00870



GPL-299

InterBioScreen

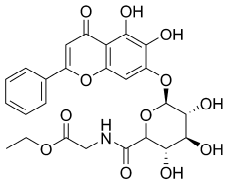
IBS_STOCK1N-45942



GPL-300

InterBioScreen

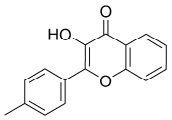
IBS_STOCK3S-83333



GPL-301

InterBioScreen

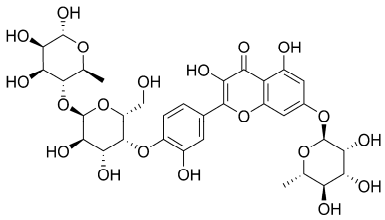
IBS_STOCK1N-06039



GPL-302

InterBioScreen

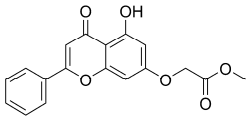
IBS_STOCK2S-88260



GPL-303

InterBioScreen

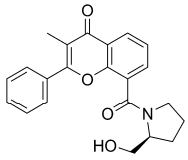
IBS_STOCK1N-68438



GPL-304

InterBioScreen

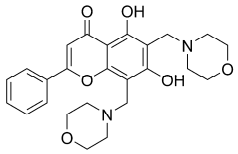
IBS_STOCK1N-13166



GPL-305

InterBioScreen

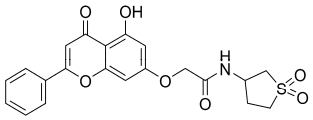
IBS_STOCK1N-69857



GPL-306

InterBioScreen

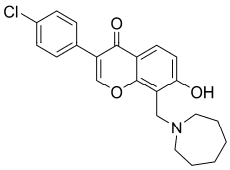
IBS_STOCK1N-29495



GPL-307

InterBioScreen

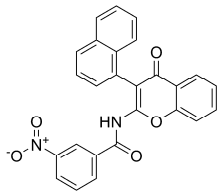
IBS_STOCK6S-14615



GPL-308

InterBioScreen

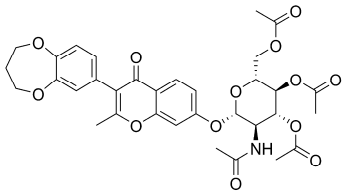
IBS_STOCK1N-26657



GPL-309

InterBioScreen

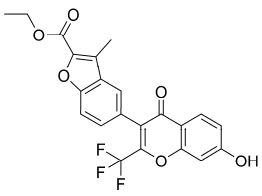
IBS_STOCK5S-66402



GPL-310

InterBioScreen

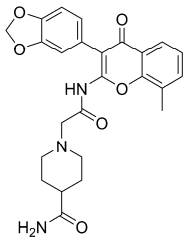
IBS_STOCK1N-25358



GPL-311

InterBioScreen

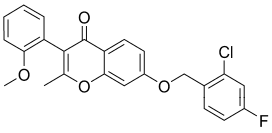
IBS_STOCK1S-17634



GPL-312

InterBioScreen

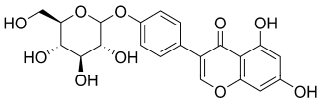
IBS_STOCK6S-09506



GPL-313

InterBioScreen

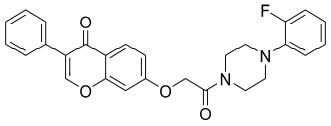
IBS_STOCK3S-87867



GPL-314

InterBioScreen

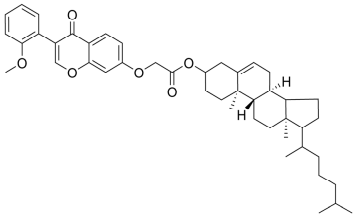
IBS_STOCK1N-02643



GPL-315

InterBioScreen

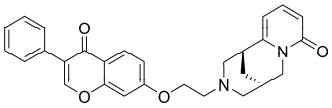
IBS_STOCK6S-38726



GPL-316

InterBioScreen

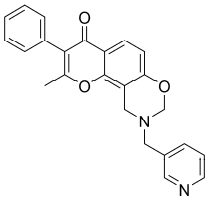
IBS_STOCK1N-34239



GPL-317

InterBioScreen

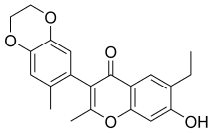
IBS_STOCK1N-68030



GPL-318

InterBioScreen

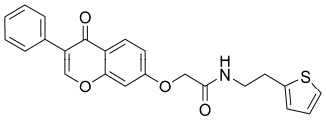
IBS_STOCK6S-36153



GPL-319

InterBioScreen

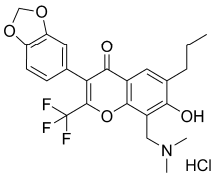
IBS_STOCK1N-02545



GPL-320

InterBioScreen

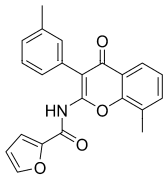
IBS_STOCK6S-40346



GPL-321

InterBioScreen

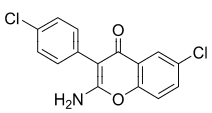
IBS_STOCK1S-16765



GPL-322

InterBioScreen

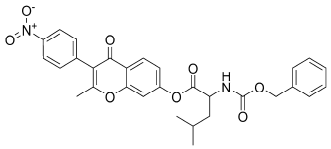
IBS_STOCK6S-02305



GPL-323

InterBioScreen

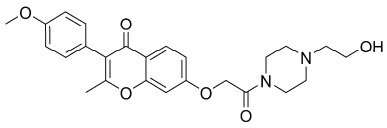
IBS_STOCK6S-18162



GPL-324

InterBioScreen

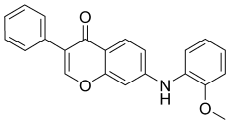
IBS_STOCK1S-36351



GPL-325

InterBioScreen

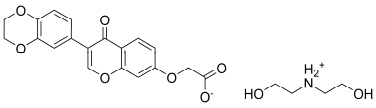
IBS_STOCK6S-42692



GPL-326

InterBioScreen

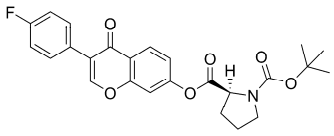
IBS_STOCK6S-12271



GPL-327

InterBioScreen

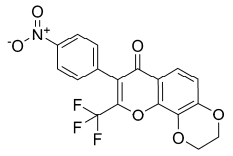
IBS_STOCK1N-04112



GPL-328

InterBioScreen

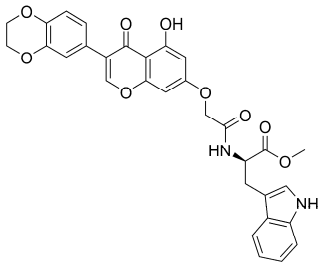
IBS_STOCK5S-39341



GPL-329

InterBioScreen

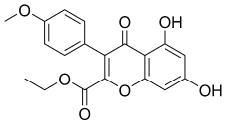
IBS_STOCK1S-18155



GPL-330

InterBioScreen

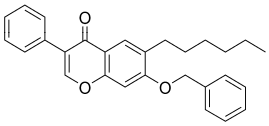
IBS_STOCK1N-01391



GPL-331

InterBioScreen

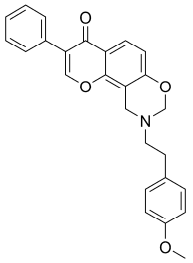
IBS_STOCK1N-19481



GPL-332

InterBioScreen

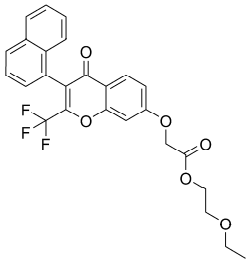
IBS_STOCK5S-43236



GPL-333

InterBioScreen

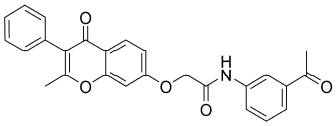
IBS_STOCK6S-34989



GPL-334

InterBioScreen

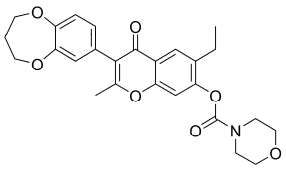
IBS_STOCK2S-61709



GPL-335

InterBioScreen

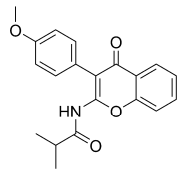
IBS_STOCK6S-37503



GPL-336

InterBioScreen

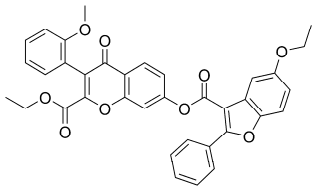
IBS_STOCK5S-35984



GPL-337

InterBioScreen

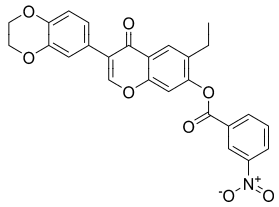
IBS_STOCK6S-26367



GPL-338

InterBioScreen

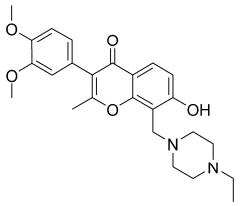
IBS_STOCK1N-29629



GPL-339

InterBioScreen

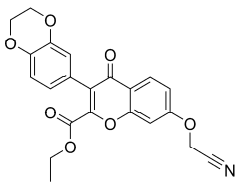
IBS_STOCK1N-40851



GPL-340

InterBioScreen

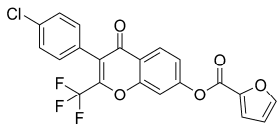
IBS_STOCK4S-56841



GPL-341

InterBioScreen

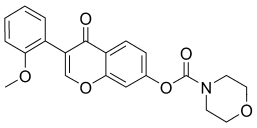
IBS_STOCK5S-49473



GPL-342

InterBioScreen

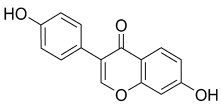
IBS_STOCK3S-05115



GPL-343

InterBioScreen

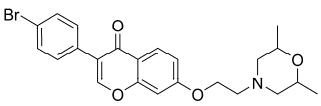
IBS_STOCK4S-54554



GPL-344

InterBioScreen

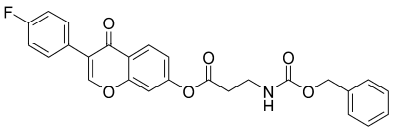
IBS_STOCK1N-31750



GPL-345

InterBioScreen

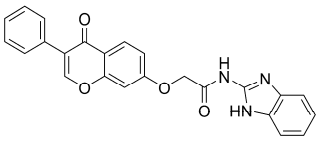
IBS_STOCK6S-17700



GPL-346

InterBioScreen

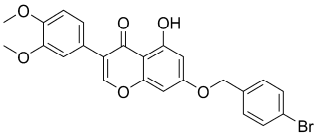
IBS_STOCK1S-36886



GPL-347

InterBioScreen

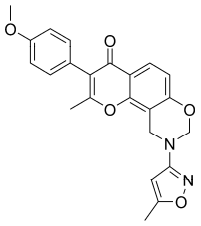
IBS_STOCK6S-40992



GPL-348

InterBioScreen

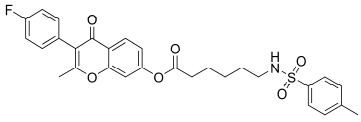
IBS_STOCK5S-46787



GPL-349

InterBioScreen

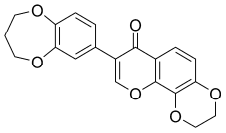
IBS_STOCK1N-70767



GPL-350

InterBioScreen

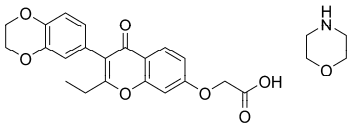
IBS_STOCK1S-32990



GPL-351

InterBioScreen

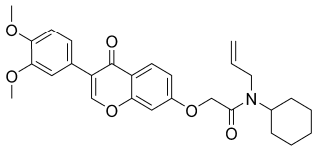
IBS_STOCK1N-03530



GPL-352

InterBioScreen

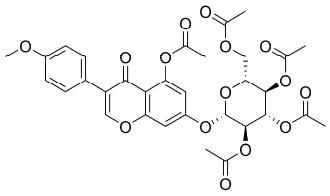
IBS_STOCK1N-05501



GPL-353

InterBioScreen

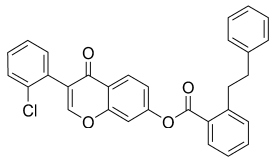
IBS_STOCK6S-43092



GPL-354

InterBioScreen

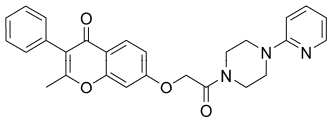
IBS_STOCK1N-23870



GPL-355

InterBioScreen

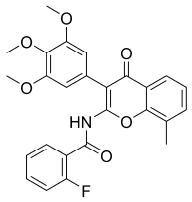
IBS_STOCK5S-47662



GPL-356

InterBioScreen

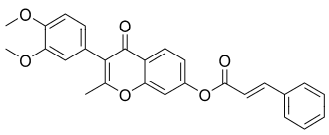
IBS_STOCK6S-36354



GPL-357

InterBioScreen

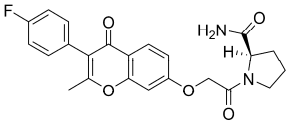
IBS_STOCK6S-13147



GPL-358

InterBioScreen

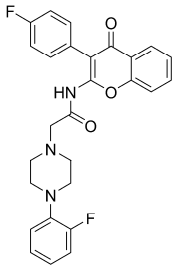
IBS_STOCK1N-39233



GPL-359

InterBioScreen

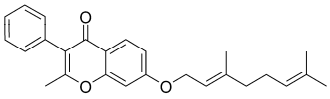
IBS_STOCK1N-04695



GPL-360

InterBioScreen

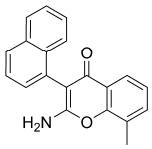
IBS_STOCK6S-06534



GPL-361

InterBioScreen

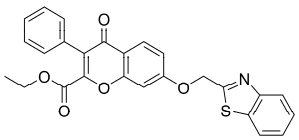
IBS_STOCK1N-40022



GPL-362

InterBioScreen

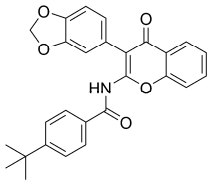
IBS_STOCK6S-17510



GPL-363

InterBioScreen

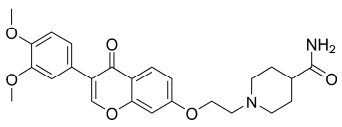
IBS_STOCK1S-24991



GPL-364

InterBioScreen

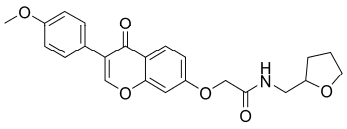
IBS_STOCK5S-78412



GPL-365

InterBioScreen

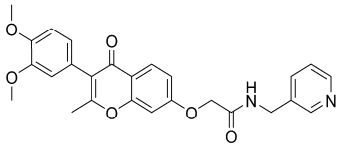
IBS_STOCK6S-29496



GPL-366

InterBioScreen

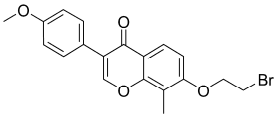
IBS_STOCK6S-33875



GPL-367

InterBioScreen

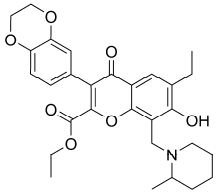
IBS_STOCK1N-70653



GPL-368

InterBioScreen

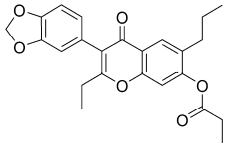
IBS_STOCK6S-17959



GPL-369

InterBioScreen

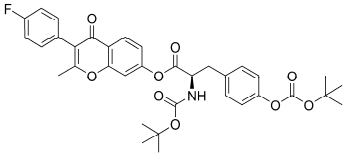
IBS_STOCK3S-50447



GPL-370

InterBioScreen

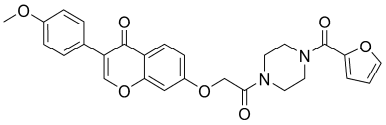
IBS_STOCK1N-07711



GPL-371

InterBioScreen

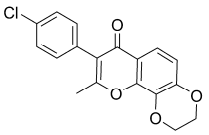
IBS_STOCK1N-06531



GPL-372

InterBioScreen

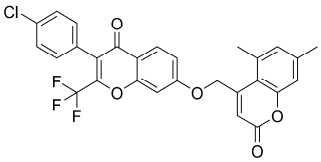
IBS_STOCK6S-43189



GPL-373

InterBioScreen

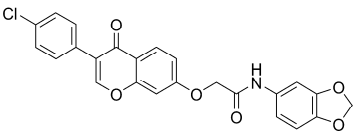
IBS_STOCK1N-02139



GPL-374

InterBioScreen

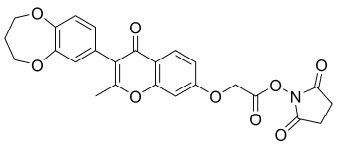
IBS_STOCK5S-51624



GPL-375

InterBioScreen

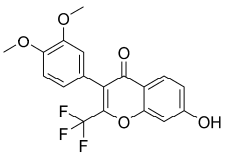
IBS_STOCK6S-37476



GPL-376

InterBioScreen

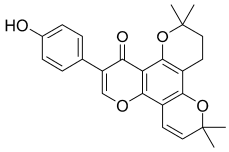
IBS_STOCK1N-00730



GPL-377

InterBioScreen

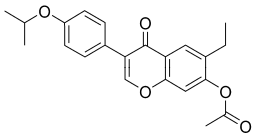
IBS_STOCK2S-56950



GPL-378

InterBioScreen

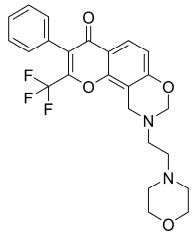
IBS_STOCK1N-12744



GPL-379

InterBioScreen

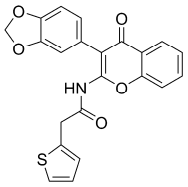
IBS_STOCK1N-01369



GPL-380

InterBioScreen

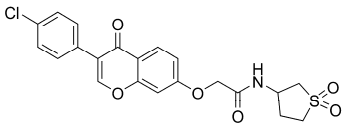
IBS_STOCK6S-42766



GPL-381

InterBioScreen

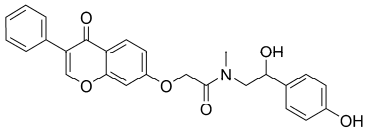
IBS_STOCK5S-88988



GPL-382

InterBioScreen

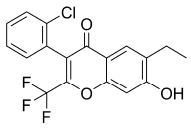
IBS_STOCK6S-37001



GPL-383

InterBioScreen

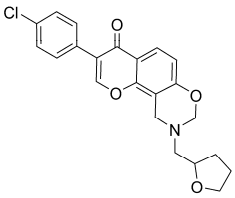
IBS_STOCK1N-70825



GPL-384

InterBioScreen

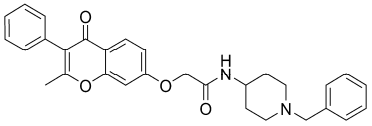
IBS_STOCK2S-94055



GPL-385

InterBioScreen

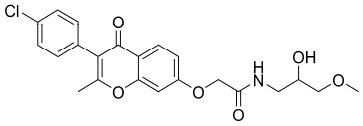
IBS_STOCK6S-41359



GPL-386

InterBioScreen

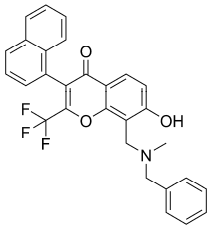
IBS_STOCK6S-41082



GPL-387

InterBioScreen

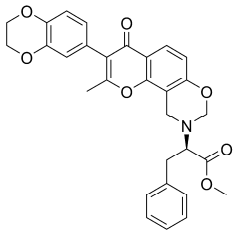
IBS_STOCK6S-36117



GPL-388

InterBioScreen

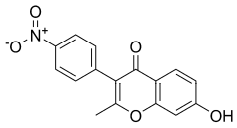
IBS_STOCK2S-80688



GPL-389

InterBioScreen

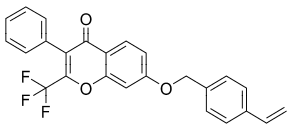
IBS_STOCK1N-05489



GPL-390

InterBioScreen

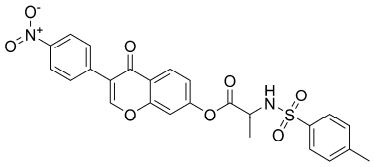
IBS_STOCK1S-15658



GPL-391

InterBioScreen

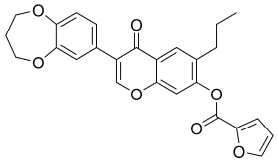
IBS_STOCK5S-99899



GPL-392

InterBioScreen

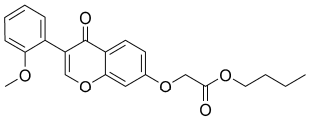
IBS_STOCK1S-31955



GPL-393

InterBioScreen

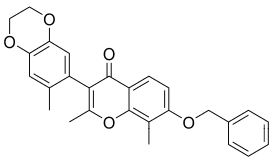
IBS_STOCK1N-37510



GPL-394

InterBioScreen

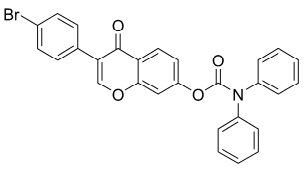
IBS_STOCK5S-44893



GPL-395

InterBioScreen

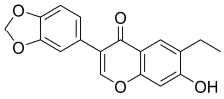
IBS_STOCK5S-56432



GPL-396

InterBioScreen

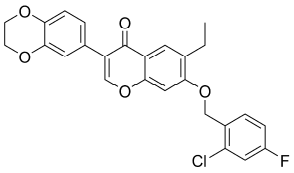
IBS_STOCK3S-48305



GPL-397

InterBioScreen

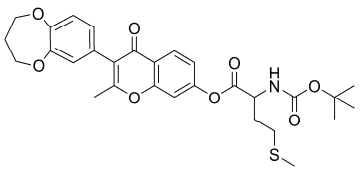
IBS_STOCK1N-04755



GPL-398

InterBioScreen

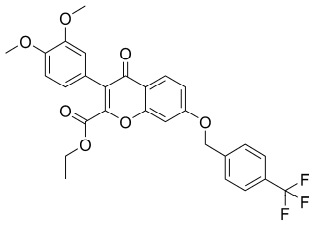
IBS_STOCK3S-49527



GPL-399

InterBioScreen

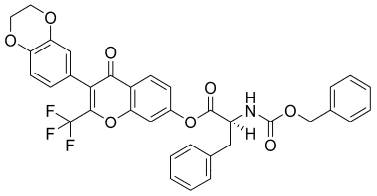
IBS_STOCK5S-38856



GPL-400

InterBioScreen

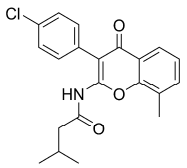
IBS_STOCK2S-90157



GPL-401

InterBioScreen

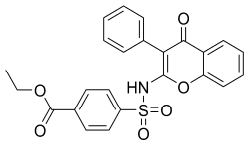
IBS_STOCK1N-06618



GPL-402

InterBioScreen

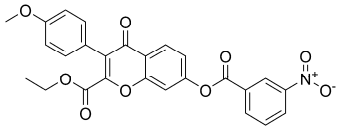
IBS_STOCK6S-04532



GPL-403

InterBioScreen

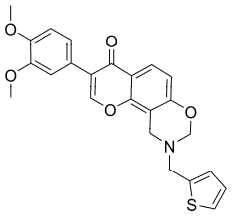
IBS_STOCK6S-36611



GPL-404

InterBioScreen

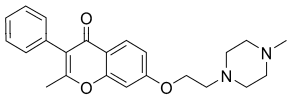
IBS_STOCK1N-33895



GPL-405

InterBioScreen

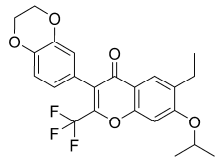
IBS_STOCK1N-70689



GPL-406

InterBioScreen

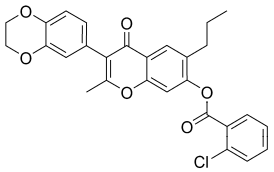
IBS_STOCK6S-18159



GPL-407

InterBioScreen

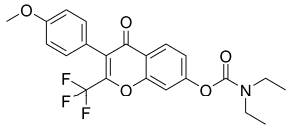
IBS_STOCK1S-19323



GPL-408

InterBioScreen

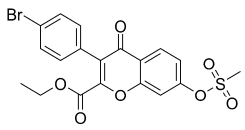
IBS_STOCK3S-47235



GPL-409

InterBioScreen

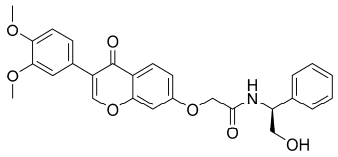
IBS_STOCK6S-00512



GPL-410

InterBioScreen

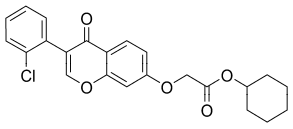
IBS_STOCK1N-33844



GPL-411

InterBioScreen

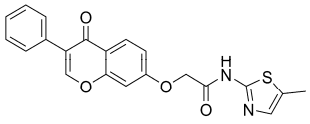
IBS_STOCK1N-70168



GPL-412

InterBioScreen

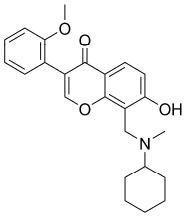
IBS_STOCK2S-63621



GPL-413

InterBioScreen

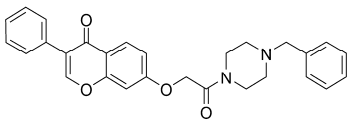
IBS_STOCK6S-43658



GPL-414

InterBioScreen

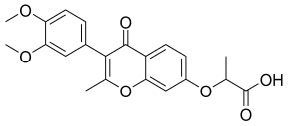
IBS_STOCK5S-52209



GPL-415

InterBioScreen

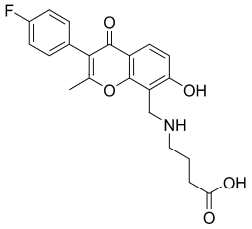
IBS_STOCK6S-37080



GPL-416

InterBioScreen

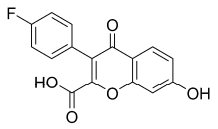
IBS_STOCK1N-40826



GPL-417

InterBioScreen

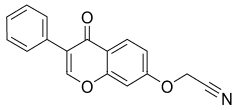
IBS_STOCK1N-09127



GPL-418

InterBioScreen

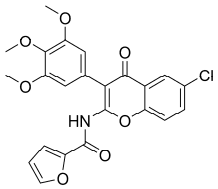
IBS_STOCK1N-04685



GPL-419

InterBioScreen

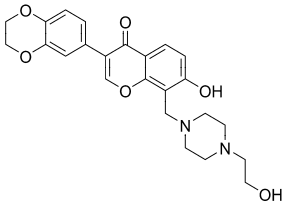
IBS_STOCK4S-39768



GPL-420

InterBioScreen

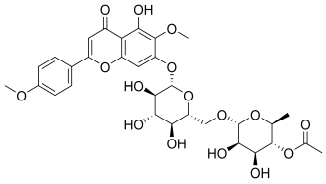
IBS_STOCK5S-67510



GPL-421

InterBioScreen

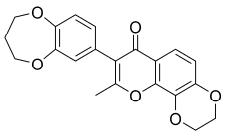
IBS_STOCK4S-68880



GPL-422

InterBioScreen

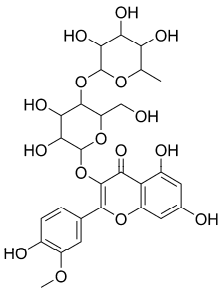
IBS_STOCK1N-01756



GPL-423

InterBioScreen

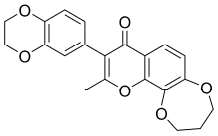
IBS_STOCK1N-03004



GPL-424

InterBioScreen

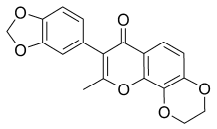
IBS_STOCK1N-04457



GPL-425

InterBioScreen

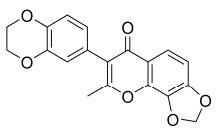
IBS_STOCK1N-05003



GPL-426

InterBioScreen

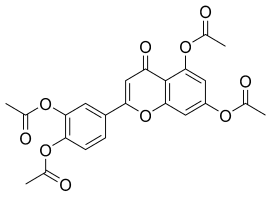
IBS_STOCK1N-05175



GPL-427

InterBioScreen

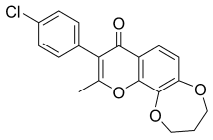
IBS_STOCK1N-05794



GPL-428

InterBioScreen

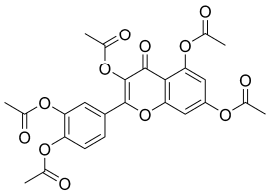
IBS_STOCK1N-05936



GPL-429

InterBioScreen

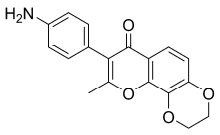
IBS_STOCK1N-06645



GPL-430

InterBioScreen

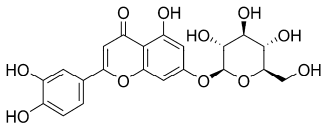
IBS_STOCK1N-08201



GPL-431

InterBioScreen

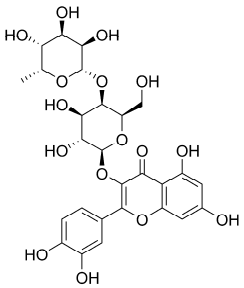
IBS_STOCK1N-08413



GPL-432

InterBioScreen

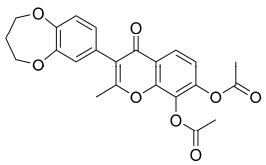
IBS_STOCK1N-08497



GPL-433

InterBioScreen

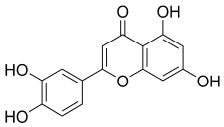
IBS_STOCK1N-09823



GPL-434

InterBioScreen

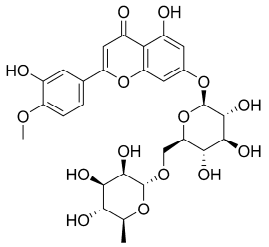
IBS_STOCK1N-11277



GPL-435

InterBioScreen

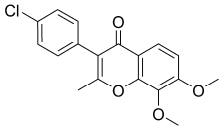
IBS_STOCK1N-14308



GPL-436

InterBioScreen

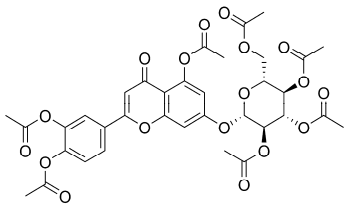
IBS_STOCK1N-14729



GPL-437

InterBioScreen

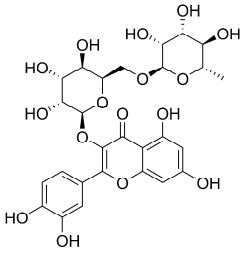
IBS_STOCK1N-15679



GPL-438

InterBioScreen

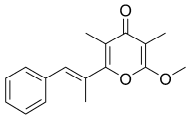
IBS_STOCK1N-17593



GPL-439

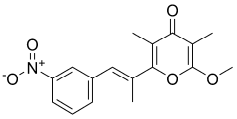
InterBioScreen

IBS_STOCK1N-23471



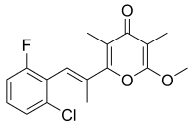
GPL-440

synthesized



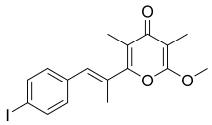
GPL-441

synthesized



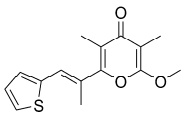
GPL-442

synthesized



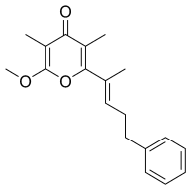
GPL-443

synthesized



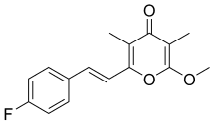
GPL-444

synthesized



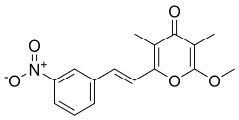
GPL-445

synthesized



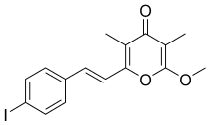
GPL-446

synthesized



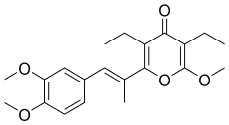
GPL-447

synthesized



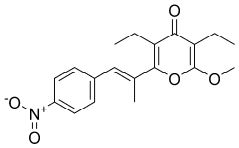
GPL-448

synthesized



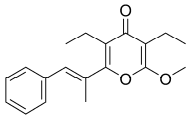
GPL-449

synthesized



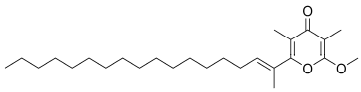
GPL-450

synthesized



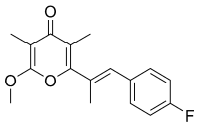
GPL-451

synthesized



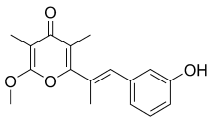
GPL-452

synthesized



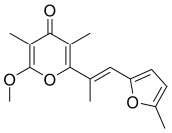
GPL-453

synthesized



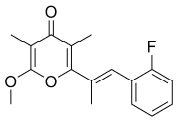
GPL-454

synthesized



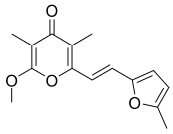
GPL-455

synthesized



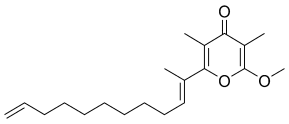
GPL-456

synthesized



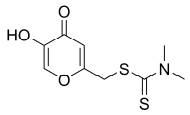
GPL-457

synthesized



GPL-458

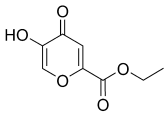
synthesized



GPL-459

Aurora

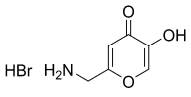
kasf-025126



GPL-460

Aurora

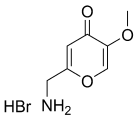
ka-10041



GPL-461

Aurora

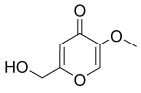
kasf-058775



GPL-462

Aurora

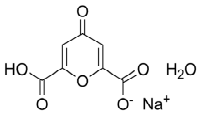
kasf-059164



GPL-463

Aurora

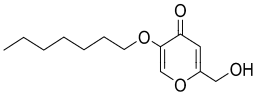
kasf-059200



GPL-464

Aurora

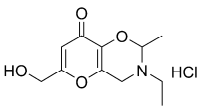
kasf-059771



GPL-465

Aurora

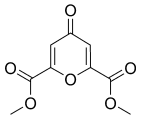
kasf-060757



GPL-466

Aurora

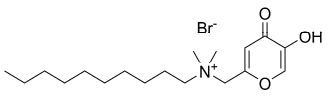
kasf-060846



GPL-467

Aurora

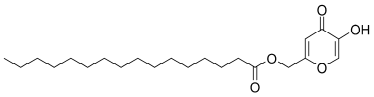
kasf-061062



GPL-468

Aurora

kasf-061630



GPL-469

Aurora

kasf-062367

GPL-470

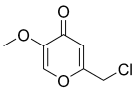
Aurora

kasf-101601

GPL-471

Aurora

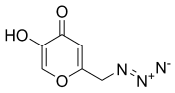
kasf-104766



GPL-472

Aurora

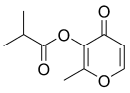
kasf-106293



GPL-473

Aurora

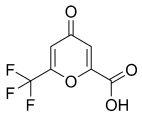
kasf-106981



GPL-474

Aurora

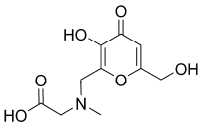
kasf-107849



GPL-475

Aurora

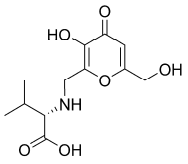
kasi-398833



GPL-476

Aurora

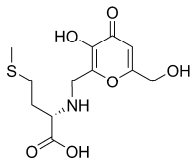
kbs-000104



GPL-477

Aurora

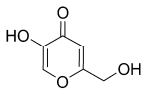
kbs-000132



GPL-478

Aurora

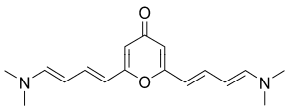
kbs-000170



GPL-479

Aurora

kbsa-0000028



GPL-480

Aurora

kbsa-0000631

GPL-481

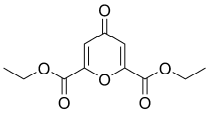
Aurora

kbsa-0000697

GPL-482

Aurora

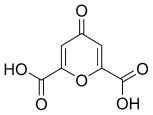
kbsa-0143980



GPL-483

Aurora

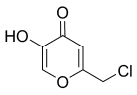
kbsb-0091512



GPL-484

Aurora

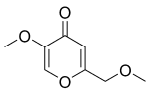
kbsenon-0000130



GPL-485

Aurora

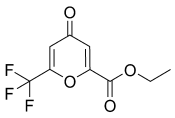
kbsenon-0009034



GPL-486

Aurora

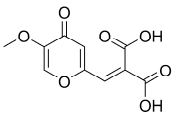
kasf-106283



GPL-487

Aurora

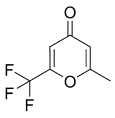
kcheb-133442



GPL-488

Aurora

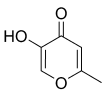
kasf-109822



GPL-489

Aurora

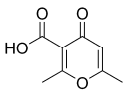
kenb-0015722



GPL-490

Aurora

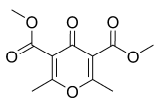
ki-034198



GPL-491

Aurora

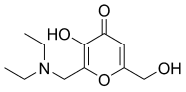
ki-035349



GPL-492

Aurora

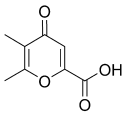
kina-0033191



GPL-493

Aurora

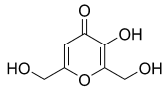
kina-0140602



GPL-494

Aurora

kmy-043452



GPL-495

Aurora

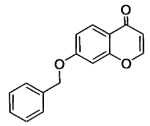
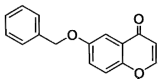
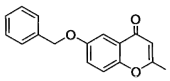
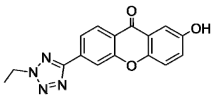
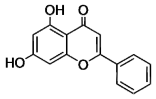
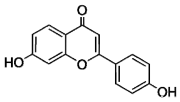
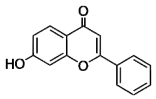
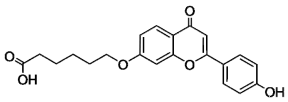
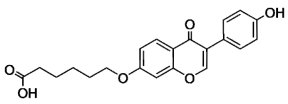
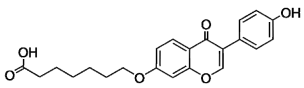
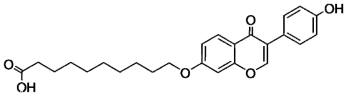
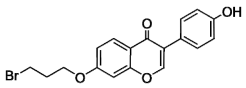
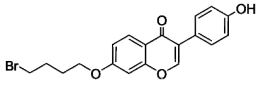
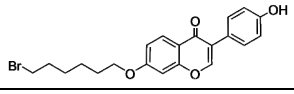
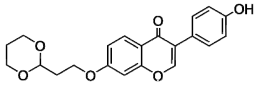
kmy-065990

GPL-496

Aurora

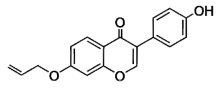
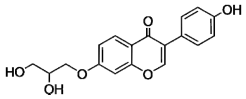
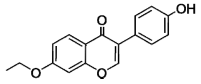
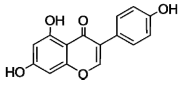
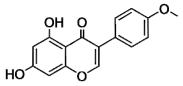
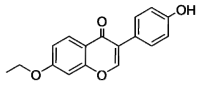
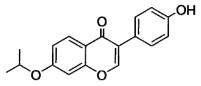
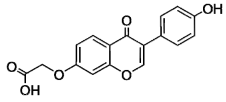
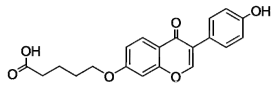
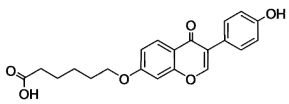
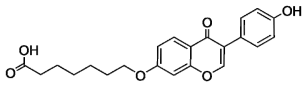
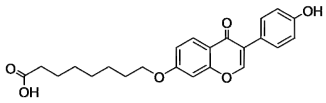
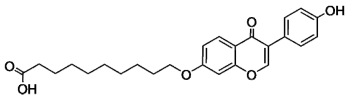
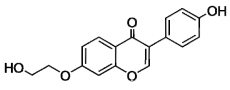
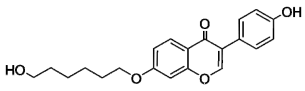
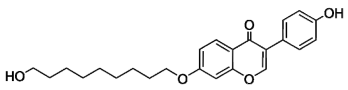
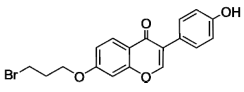
kph-118129

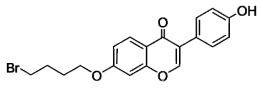
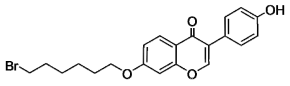
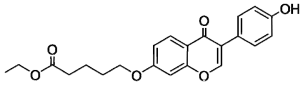
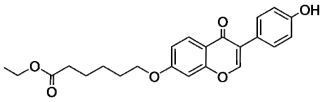
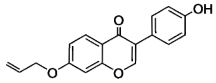
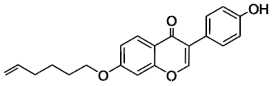
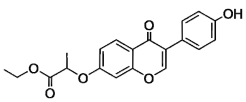
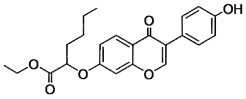
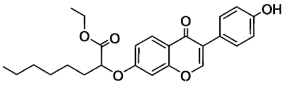
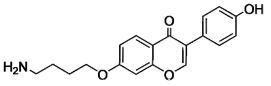
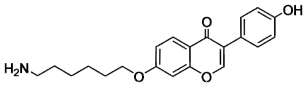
Attachment 4: MAO Inhibitors from WOMBAT with $IC_{50}/K_i \leq 10$ µM.[6]

Number	Structures	IC_{50} / K_i [µM][7]		Reference
		MAO A	MAO B	
1		--	1.05	[307]
2		6.76	0.56	[307]
		6.76	0.56	[308]
3		--	0.13	[307]
		--	0.13	[308]
4		2.00	0.10	[309]
5		1.60		[310]
6		0.03		[311]
7		0.12		[311]
8		7.00		[311]
9		3.00		[310]
10		2.10		[310]
11		10.00		[310]
12		0.15		[310]
13		4.00		[310]
14		2.00		[310]
15		7.00		[310]

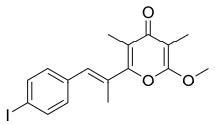
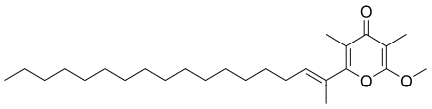
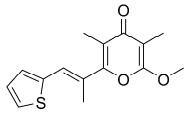
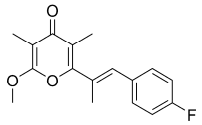
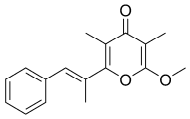
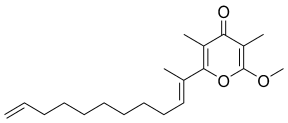
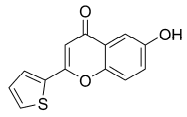
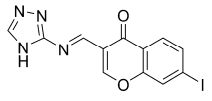
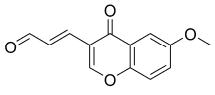
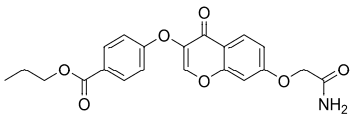
[6] "MAO" denotes inhibition of total MAO activity.

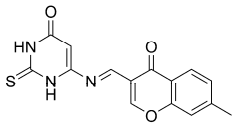
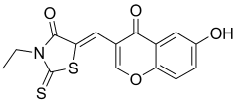
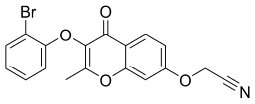
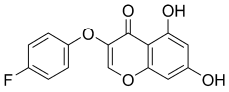
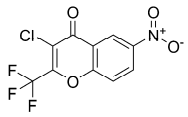
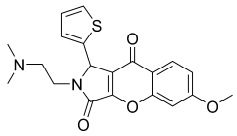
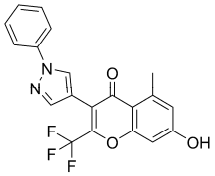
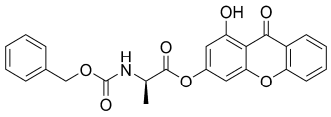
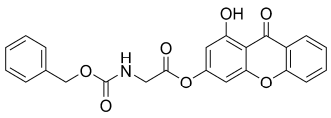
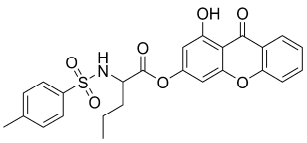
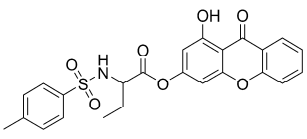
[7] Values written across both columns denote have been measured as inhibition of the total MAO activity.

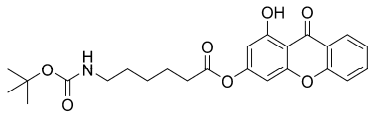
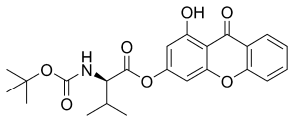
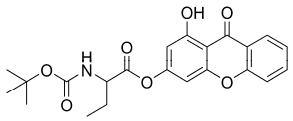
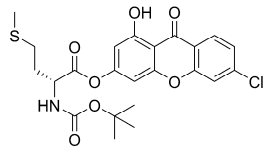
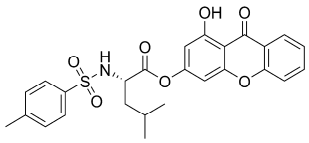
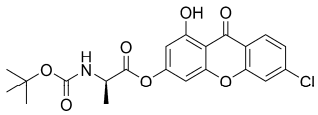
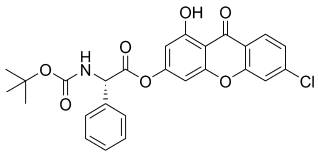
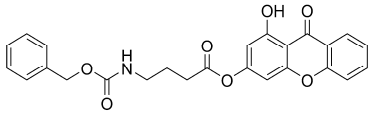
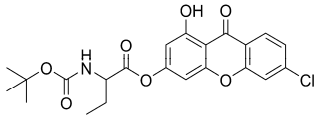
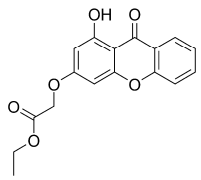
Number	Structures	IC_{50} / K_i [µM][7]		Reference
		MAO A	MAO B	
16		0.45		[310]
17		1.70		[310]
18		0.30		[310]
19		0.90		[310]
20		0.40		[310]
21		0.30		[311]
22		0.20		[311]
23		9.00		[311]
24		5.00		[311]
25		4.00		[311]
26		2.10		[311]
27		6.50		[311]
28		10.00		[311]
29		1.50		[311]
30		9.00		[311]
31		9.00		[311]
32		0.15		[311]

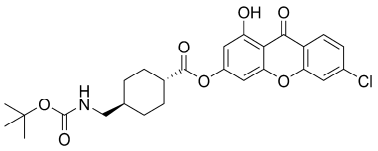
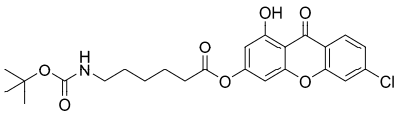
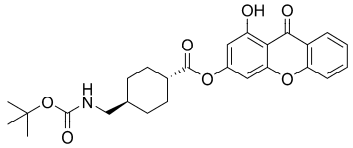
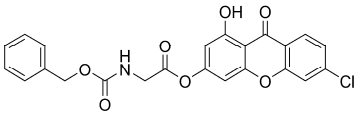
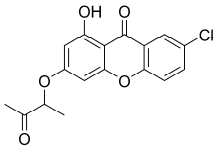
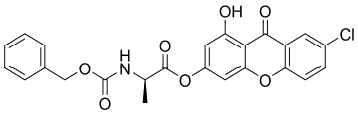
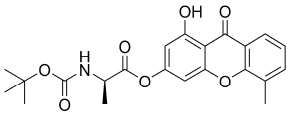
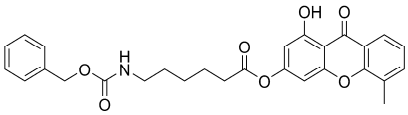
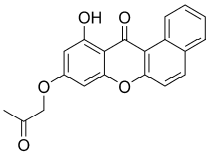
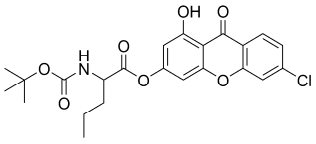
Number	Structures	IC_{50} / K_i [μM][7]		Reference
		MAO A	MAO B	
33		4.00		[311]
34		2.00		[311]
35		9.00		[311]
36		5.30		[311]
37		0.45		[311]
38		5.00		[311]
39		9.00		[311]
40		9.00		[311]
41		9.00		[311]
42		9.00		[311]
43		9.00		[311]

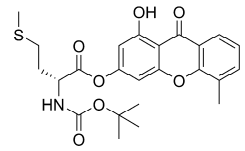
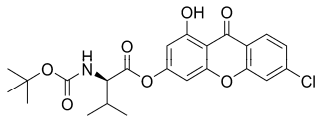
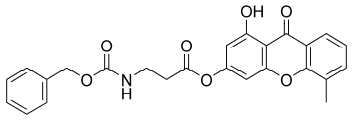
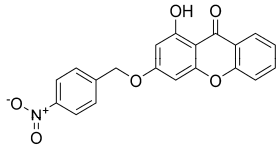
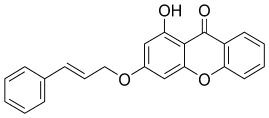
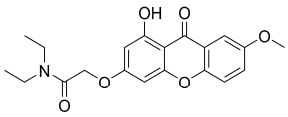
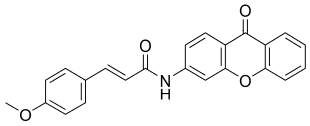
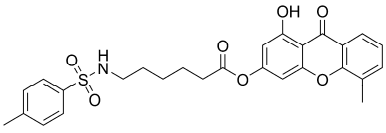
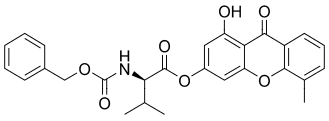
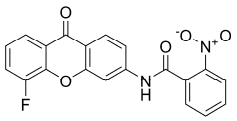
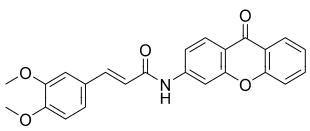
Attachment 5: Inhibitors of monoamine oxidase from the γ-pyrone library. The IC_{50} values for both enzymes were determined. If no value is given, the IC_{50} is larger than 100 µM. The numbers refer to the numbering of the complete library in Attachment 3.

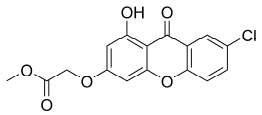
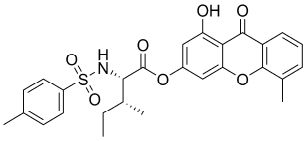
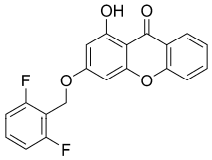
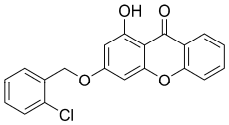
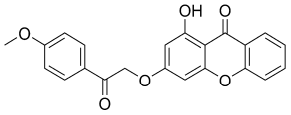
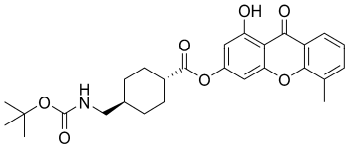
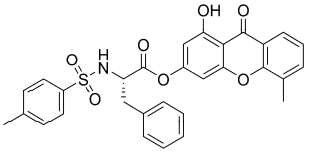
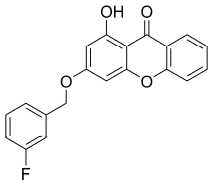
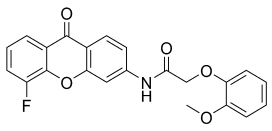
No.	Structures	IC_{50} [µM]	
		MAOA	MAOB
448		3.19 ± 0.23	4.02 ± 0.23
452		12.5 ± 1.0	
444		16.3 ± 0.5	
453		18.1 ± 1.5	
451		18.2 ± 1.1	
458			4.88 ± 0.18
4		0.94 ± 0.03	
18		1.31 ± 0.04	
76		3.96 ± 0.19	
91		6.14 ± 0.40	

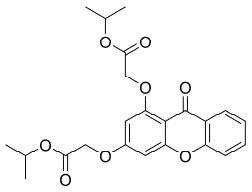
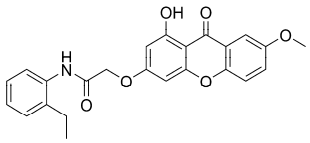
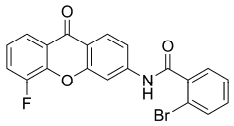
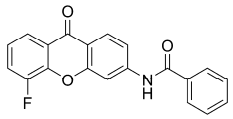
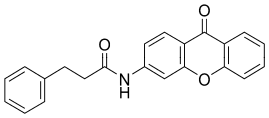
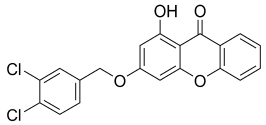
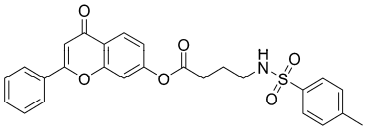
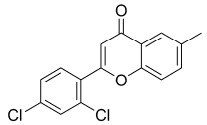
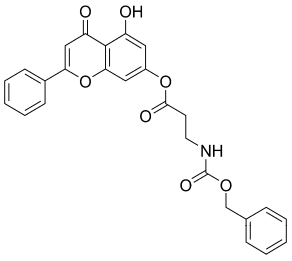
No.	Structures	IC_{50} [µM]	
		MAOA	MAOB
32		7.25 ± 1.01	1.14 ± 0.09
33		18.5 ± 1.1	
86		20.1 ± 2.8	0.715 ± 0.058
44		34.1 ± 4.2	1.41 ± 0.10
87		40.1 ± 2.9	
75			0.342 ± 0.170
96			1.43 ± 0.16
141		0.231 ± 0.014	1.92 ± 0.10
106		0.438 ± 0.028	1.54 ± 0.24
196		0.532 ± 0.045	4.04 ± 0.41
100		0.565 ± 0.032	3.94 ± 0.16

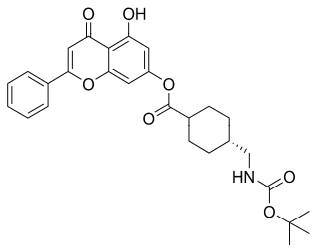
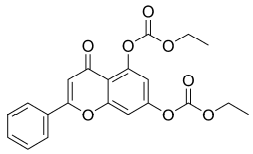
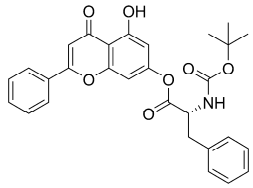
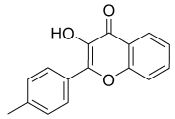
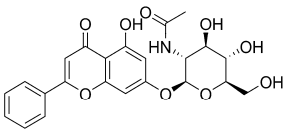
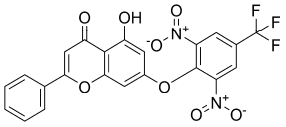
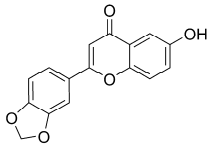
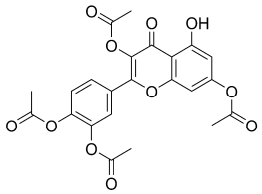
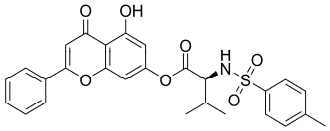
No.	Structures	IC_{50} [µM]	
		MAOA	MAOB
144		0.624 ± 0.081	
188		0.624 ± 0.034	
130		0.713 ± 0.038	
145		0.826 ± 0.040	1.50 ± 0.08
153		0.850 ± 0.018	4.95 ± 0.14
150		0.959 ± 0.103	1.42 ± 0.15
120		0.972 ± 0.056	1.34 ± 0.07
157		1.05 ± 0.07	
189		1.06 ± 0.09	
123		1.07 ± 0.08	3.81 ± 1.19

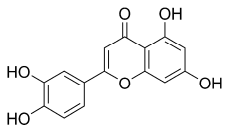
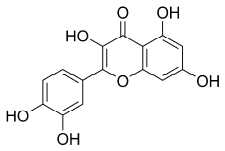
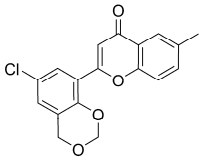
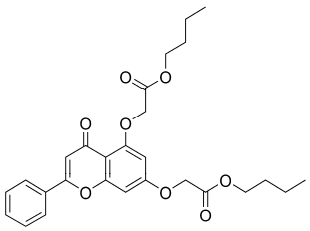
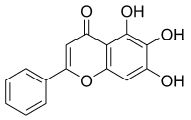
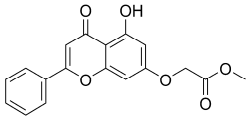
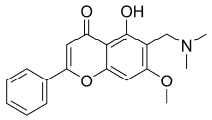
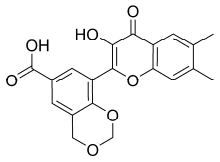
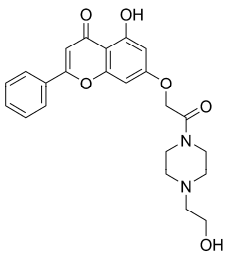
No.	Structures	IC_{50} [µM]	
		MAOA	MAOB
166		1.31 ± 0.24	
169		1.41 ± 0.14	
176		1.41 ± 0.06	1.75 ± 0.33
168		2.09 ± 0.17	0.562 ± 0.088
191		2.12 ± 0.18	
137		2.42 ± 0.24	1.34 ± 0.08
160		2.63 ± 0.35	15.6 ± 2.1
165		2.73 ± 0.67	
194		2.85 ± 0.26	2.80 ± 0.18
108		3.42 ± 0.26	

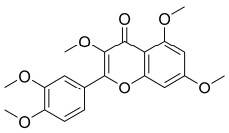
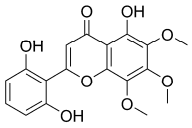
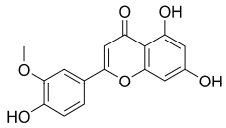
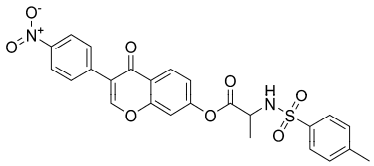
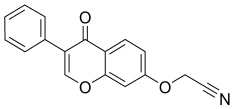
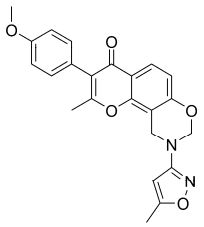
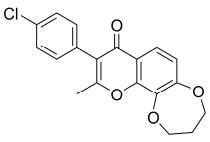
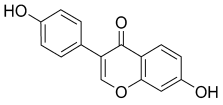
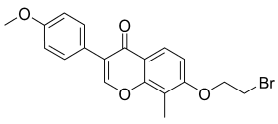
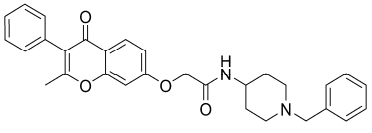
No.	Structures	IC_{50} [µM]	
		MAOA	MAOB
118		3.49 ± 0.13	
181		4.33 ± 0.32	4.87 ± 0.44
174		5.04 ± 0.75	0.218 ± 0.035
197		5.08 ± 0.48	
182		5.16 ± 0.51	
171		5.86 ± 0.35	
185		7.44 ± 0.49	
132		6.99 ± 0.36	
117		9.81 ± 0.60	
101		10.4 ± 0.5	
102		12.5 ± 0.8	

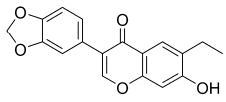
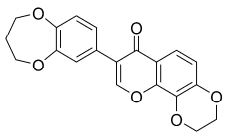
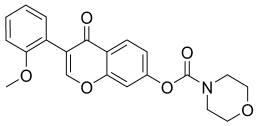
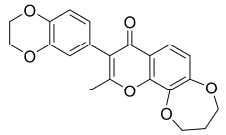
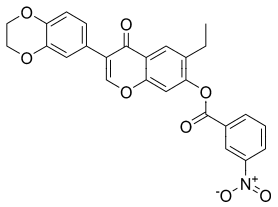
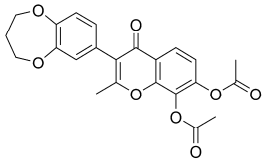
No.	Structures	IC_{50} [µM]	
		MAOA	MAOB
113		14.5 ± 1.0	14.0 ± 1.6
128		15.3 ± 1.6	9.47 ± 1.89
164		18.1 ± 1.2	
112		19.9 ± 0.9	0.143 ± 0.008
458		21.3 ± 1.3	
121		22.4 ± 1.5	
146		22.6 ± 1.3	
175		24.6 ± 1.3	1.92 ± 0.16
179		32.8 ± 3.3	

No.	Structures	IC_{50} [μM]	
		MAOA	MAOB
159		83.6 ± 8.9	
110			0.051 ± 0.002
154			0.740 ± 0.054
131			5.42 ± 0.87
151			12.0 ± 1.3
142			17.4 ± 1.7
298		0.886 ± 0.053	
300		0.945 ± 0.068	
235		0.953 ± 0.047	

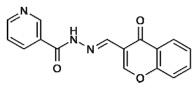
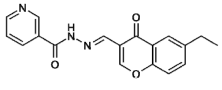
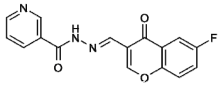
No.	Structures	IC_{50} [μM]	
		MAOA	MAOB
289		1.31 ± 0.07	
247		1.99 ± 0.10	
211		2.45 ± 0.06	
302		3.37 ± 0.45	
217		3.91 ± 0.30	2.48 ± 0.36
213		4.17 ± 0.13	
274		4.29 ± 0.48	0.572 ± 0.059
205		4.89 ± 0.36	
276		7.57 ± 0.73	

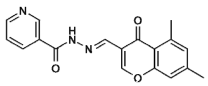
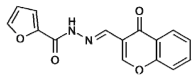
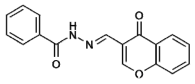
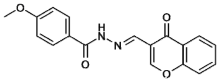
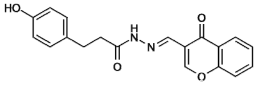
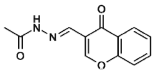
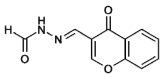
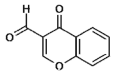
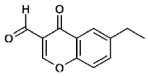
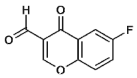
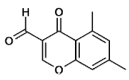
No.	Structures	IC_{50} [μM]	
		MAOA	MAOB
435		7.72 ± 0.51	20.6 ± 2.3
206		8.01 ± 0.41	
292		10.3 ± 1.3	
293		10.8 ± 1.1	
233		10.4 ± 0.3	22.1 ± 2.0
304		11.1 ± 0.9	
236		11.9 ± 0.7	
204		28.5 ± 2.6	
246		29.4 ± 2.3	

No.	Structures	IC_{50} [μM]	
		MAOA	MAOB
281		35.1 ± 1.8	
254		37.2 ± 5.5	26.7 ± 3.6
222			15.1 ± 1.7
392		1.79 ± 0.03	
419		2.33 ± 0.24	
349		2.86 ± 0.12	
429		6.39 ± 0.60	7.51 ± 1.26
344		9.88 ± 0.79	
368		10.2 ± 1.03	0.865 ± 0.051
386		39.2 ± 3.3	

No.	Structures	IC_{50} [µM]	
		MAOA	MAOB
397			0.304 ± 0.031
351			0.544 ± 0.057
343			1.470 ± 0.177
425			5.01 ± 0.85
339			14.5 ± 0.9
434			34.8 ± 2.2

Attachment 6: Chromone-based STAT inhibitors as published by T. Berg *et al*.[110]

Number	Structures	App. IC_{50} [µM]		
		STAT5b	STAT3	STAT1
1		47 ± 17	> 500	> 500
2		53 ± 32	54 ± 8	52 ± 1
3		79 ± 20	159 ± 19	396 ± 84

Number	Structures	App. IC_{50} [µM]		
		STAT5b	STAT3	STAT1
4		217 ± 45	107 ± 9	162 ± 36
5		56 ± 10	> 500	> 500
6		53 ± 24	241 ± 55	>500
7		64 ± 26	176 ± 22	351 ± 23
8		90 ± 30	> 500	> 500
9		86 ± 27	242 ± 43	> 500
10		92 ± 13	343 ± 14	> 500
11		15 ± 1	54 ± 7	69 ± 5
12		11 ± 2	20 ± 2	34 ± 5
13		22 ± 4	41 ± 11	64 ± 8
14		51 ± 10	n.d.	n.d.

Attachment 7: The list of 33 PSSC clusters with less than 50% of their members predicted by SCOP. Individual clusters are divided by bold lines; the first entry of each cluster (without data for alignment length and sequence identity) is the cluster center.

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
11as	25.3	0			6.3.1.1	Aspartate--ammonia ligase (EC 6.3.1.1) (Asparagine synthetase A) [Gene: asnA or b3744 or JW3722] - Escherichia coli (strain K12)
12as	22.9	0.3	115	100	6.3.1.1	Aspartate--ammonia ligase (EC 6.3.1.1) (Asparagine synthetase A) [Gene: asnA or b3744 or JW3722] - Escherichia coli (strain K12)
1e24	11.4	1.7	95	16	6.1.1.6	Lysyl-tRNA synthetase, heat inducible (EC 6.1.1.6) (Lysine--tRNA ligase) (LysRS) [Gene: lysU or b4129 or JW4090] - Escherichia coli (strain K12)
1e1t	11.1	1.6	96	16	6.1.1.6	Lysyl-tRNA synthetase, heat inducible (EC 6.1.1.6) (Lysine--tRNA ligase) (LysRS) [Gene: lysU or b4129 or JW4090] - Escherichia coli (strain K12)
1lkh	11.1	2	99	16	not found	Asparaginyl-tRNA synthetase (EC 6.1.1.22) (Asparagine--tRNA ligase) (AsnRS) [Gene: asnS or tss or STM1000] - Salmonella typhimurium
1lyl	11	2	100	14	6.1.1.6	Lysyl-tRNA synthetase, heat inducible (EC 6.1.1.6) (Lysine--tRNA ligase) (LysRS) [Gene: lysU or b4129 or JW4090] - Escherichia coli (strain K12)
1asz	10.8	1.8	95	24	6.1.1.12	Aspartyl-tRNA synthetase, cytoplasmic (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS) [Gene: DPS1 or APS or APS1 or YLL018C or L1295] - Saccharomyces cerevisiae (Baker's yeast)
1b8a	10.8	1.8	93	26	6.1.1.12	Aspartyl-tRNA synthetase (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS) [Gene: aspS or TK0492] - Pyrococcus kodakaraensis (Thermococcus kodakaraensis)
1e1o	10.3	1.8	94	15	6.1.1.6	Lysyl-tRNA synthetase, heat inducible (EC 6.1.1.6) (Lysine--tRNA ligase) (LysRS) [Gene: lysU or b4129 or JW4090] - Escherichia coli (strain K12)
1bbu	10	1.8	94	15	6.1.1.6	Lysyl-tRNA synthetase (EC 6.1.1.6) (Lysine--tRNA ligase) (LysRS) [Gene: lysS or asuD or herC or b2890 or JW2858] - Escherichia coli (strain K12)
1eov	9.7	2.3	91	24	6.1.1.12	Aspartyl-tRNA synthetase, cytoplasmic (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS) [Gene: DPS1 or APS or APS1 or YLL018C or L1295] - Saccharomyces cerevisiae (Baker's yeast)
1asy	9.5	2.2	91	25	6.1.1.12	Aspartyl-tRNA synthetase, cytoplasmic (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS) [Gene: DPS1 or APS or APS1 or YLL018C or L1295] - Saccharomyces cerevisiae (Baker's yeast)
1b70	8.8	2.6	95	15	6.1.1.20	Phenylalanyl-tRNA synthetase alpha chain (EC 6.1.1.20) (Phenylalanine--tRNA ligase alpha chain) (PheRS) [Gene: pheS] - Thermus thermophilus
1eiy	8.1	2.4	90	16	6.1.1.20	Phenylalanyl-tRNA synthetase alpha chain (EC 6.1.1.20) (Phenylalanine--tRNA ligase alpha chain) (PheRS) [Gene: pheS or TTHA1958] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
2akw	8.1	2.5	98	14	6.1.1.20	Phenylalanyl-tRNA synthetase alpha chain (EC 6.1.1.20) (Phenylalanine--tRNA ligase alpha chain) (PheRS) [Gene: pheS] - Thermus thermophilus
2amc	7.8	2.9	98	14	6.1.1.20	Phenylalanyl-tRNA synthetase alpha chain (EC 6.1.1.20) (Phenylalanine--tRNA ligase alpha chain) (PheRS) [Gene: pheS] - Thermus thermophilus

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1g51	7.3	1.7	72	22	6.1.1.12	Aspartyl-tRNA synthetase (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS) [Gene: aspS] - Thermus thermophilus
1n9w	4.3	2	67	27	not found	Crystal structure of the non-discriminating and archaeal-type aspartyl-tRNA synthetase from Thermus thermophilus
1a0i	24.9	0			6.5.1.1	DNA ligase (EC 6.5.1.1) (Polydeoxyribonucleotide synthase [ATP]) [Gene: 1.3] - Bacteriophage T7
1fvi	14.1	1.9	104	23	not found	CRYSTAL STRUCTURE OF CHLORELLA VIRUS DNA LIGASE-ADENYLATE
1p8l	12.4	2	97	26	not found	New Crystal Structure of Chlorella Virus DNA Ligase-Adenylate
1b04	9.5	2.2	91	16	6.5.1.2	DNA ligase (EC 6.5.1.2) (Polydeoxyribonucleotide synthase [NAD+]) [Gene: ligA or lig] - Bacillus stearothermophilus (Geobacillus stearothermophilus)
1tae	9.2	3.2	103	18	not found	DNA ligase (EC 6.5.1.2) (Polydeoxyribonucleotide synthase [NAD+]) [Gene: ligA or EF_0722] - Enterococcus faecalis (Streptococcus faecalis)
1dgt	8.8	3.2	97	16	not found	DNA ligase (EC 6.5.1.2) (Polydeoxyribonucleotide synthase [NAD+]) (Tfi DNA ligase) [Gene: ligA] - Thermus filiformis
1v9p	8.5	2.8	93	20	6.5.1.2	DNA ligase (EC 6.5.1.2) (Polydeoxyribonucleotide synthase [NAD+]) (Tfi DNA ligase) [Gene: ligA] - Thermus filiformis
1dgs	8.5	2.8	86	17	6.5.1.2	DNA ligase (EC 6.5.1.2) (Polydeoxyribonucleotide synthase [NAD+]) (Tfi DNA ligase) [Gene: ligA] - Thermus filiformis
1zau	8.1	2.3	86	19	6.5.1.2	DNA ligase (EC 6.5.1.2) (Polydeoxyribonucleotide synthase [NAD+]) [Gene: ligA or lig or Rv3014c or MT3094 or MTV012.28c] - Mycobacterium tuberculosis
1tae	8.1	2.7	94	19	not found	DNA ligase (EC 6.5.1.2) (Polydeoxyribonucleotide synthase [NAD+]) [Gene: ligA or EF_0722] - Enterococcus faecalis (Streptococcus faecalis)
1tae	7.5	2.5	85	20	not found	DNA ligase (EC 6.5.1.2) (Polydeoxyribonucleotide synthase [NAD+]) [Gene: ligA or EF_0722] - Enterococcus faecalis (Streptococcus faecalis)
1a2o	33.4	0			3.1.1.61	Chemotaxis response regulator protein-glutamate methylesterase (EC 3.1.1.61) [Gene: cheB or STM1917] - Salmonella typhimurium
1egh	5.8	2.5	82	10	4.2.3.3	Methylglyoxal synthase (EC 4.2.3.3) (MGS) [Gene: mgsA or yccG or b0963 or JW5129] - Escherichia coli (strain K12)
1ik4	5.7	2.4	81	10	4.2.3.3	Methylglyoxal synthase (EC 4.2.3.3) (MGS) [Gene: mgsA or yccG or b0963 or JW5129] - Escherichia coli (strain K12)
1vmd	5.6	2.8	85	12	4.2.3.3	Methylglyoxal synthase (EC 4.2.3.3) (MGS) [Gene: mgsA or TM_1185] - Thermotoga maritima
1b93	5.5	2.3	78	9	4.2.3.3	Methylglyoxal synthase (EC 4.2.3.3) (MGS) [Gene: mgsA or yccG or b0963 or JW5129] - Escherichia coli (strain K12)
1s89	5.4	2.5	80	9	4.2.3.3	Methylglyoxal synthase (EC 4.2.3.3) (MGS) [Gene: mgsA or yccG or b0963 or JW5129] - Escherichia coli (strain K12)
1s8a	5.4	2.5	80	9	4.2.3.3	Methylglyoxal synthase (EC 4.2.3.3) (MGS) [Gene: mgsA or yccG or b0963 or JW5129] - Escherichia coli (strain K12)
1wo8	5.3	2.5	80	14	not found	Methylglyoxal synthase (EC 4.2.3.3) (MGS) [Gene: mgsA or TTHA1794] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1a41	16.5	0			5.99.1.2	DNA topoisomerase 1 (EC 5.99.1.2) (DNA topoisomerase I) (Late protein H7) [Gene: TOP1 or VACWR104 or H7] - Vaccinia virus (strain Western Reserve / WR) (VACV)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1fag	4.1	3.1	66	5	1.14.14.1	Bifunctional P-450/NADPH-P450 reductase (Cytochrome P450(BM-3)) (P450BM-3) [Includes: Cytochrome P450 102 (EC 1.14.14.1); NADPH--cytochrome P450 reductase (EC 1.6.2.4)] [Gene: CYP102A1 or cyp102] - Bacillus megaterium
1a49	32.4	0			2.7.1.40	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40) (Pyruvate kinase muscle isozyme) [Gene: PKM2] - Oryctolagus cuniculus (Rabbit)
1a5u	31.6	0.1	163	100	2.7.1.40	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40) (Pyruvate kinase muscle isozyme) [Gene: PKM2] - Oryctolagus cuniculus (Rabbit)
1zjh	27.9	0.5	162	99	2.7.1.40	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40) (Pyruvate kinase muscle isozyme) (Pyruvate kinase 2/3) (Cytosolic thyroid hormone-binding protein) (CTHBP) (THBP1) [Gene: PKM2 or PK2 or PK3 or PKM] - Homo sapiens (Human)
1f3w	27.7	0.9	161	100	2.7.1.40	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40) (Pyruvate kinase muscle isozyme) [Gene: PKM2] - Oryctolagus cuniculus (Rabbit)
1f3x	27.6	0.9	161	100	2.7.1.40	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40) (Pyruvate kinase muscle isozyme) [Gene: PKM2] - Oryctolagus cuniculus (Rabbit)
1t5a	27.5	0.9	161	99	2.7.1.40	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40) (Pyruvate kinase muscle isozyme) (Pyruvate kinase 2/3) (Cytosolic thyroid hormone-binding protein) (CTHBP) (THBP1) [Gene: PKM2 or PK2 or PK3 or PKM] - Homo sapiens (Human)
1aqf	27.4	0.9	162	100	2.7.1.40	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40) (Pyruvate kinase muscle isozyme) [Gene: PKM2] - Oryctolagus cuniculus (Rabbit)
1liy	27.3	1.1	162	86	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1lix	27	1	162	86	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1liw	27	1.1	162	85	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1liw	27	1.1	162	85	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1liu	27	1.1	161	86	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1liy	26.8	1.2	162	86	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1liw	26.7	1	161	86	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1lix	26.5	1.2	160	86	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1lix	26.4	0.8	156	87	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1liw	26.4	1	161	85	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1e0u	26.3	0.8	162	73	2.7.1.40	Pyruvate kinase I (EC 2.7.1.40) (PK-1) [Gene: pykF or b1676 or JW1666] - Escherichia coli (strain K12)
1liy	26	1	160	87	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1liy	25.6	1.1	160	86	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1lix	25.3	1.2	160	86	2.7.1.40	Pyruvate kinase isozymes R/L (EC 2.7.1.40) (R-type/L-type pyruvate kinase) (Red cell/liver pyruvate kinase) (Pyruvate kinase 1) [Gene: PKLR or PK1 or PKL] - Homo sapiens (Human)
1a5u	25.1	1.4	162	99	2.7.1.40	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40) (Pyruvate kinase muscle isozyme) [Gene: PKM2] - Oryctolagus cuniculus (Rabbit)
1pky	24.5	0.8	162	73	2.7.1.40	Pyruvate kinase I (EC 2.7.1.40) (PK-1) [Gene: pykF or b1676 or JW1666] - Escherichia coli (strain K12)
1e0t	24.1	0.9	148	72	2.7.1.40	Pyruvate kinase I (EC 2.7.1.40) (PK-1) [Gene: pykF or b1676 or JW1666] - Escherichia coli (strain K12)
1pkl	23.8	1	149	77	2.7.1.40	Pyruvate kinase (EC 2.7.1.40) (PK) [Gene: PYK] - Leishmania mexicana
1pkm	12.2	1.1	83	98	2.7.1.40	Pyruvate kinase isozyme M1 (EC 2.7.1.40) (Pyruvate kinase muscle isozyme) [Gene: PKM2] - Felis silvestris catus (Cat)
1dxe	11.1	2.2	108	19	4.1.2.20	2-dehydro-3-deoxyglucarate aldolase (EC 4.1.2.20) (2-keto-3-deoxyglucarate aldolase) (2-dehydro-3-deoxygalactarate aldolase) (DDG aldolase) (5-keto-4-deoxy-D-glucarate aldolase) (KDGlucA) [Gene: garL or yhaF or b3126 or JW3095] - Escherichia coli (strain K12)
1dxf	11	2.2	108	19	4.1.2.20	2-dehydro-3-deoxyglucarate aldolase (EC 4.1.2.20) (2-keto-3-deoxyglucarate aldolase) (2-dehydro-3-deoxygalactarate aldolase) (DDG aldolase) (5-keto-4-deoxy-D-glucarate aldolase) (KDGlucA) [Gene: garL or yhaF or b3126 or JW3095] - Escherichia coli (strain K12)
1rpx	10.8	2.6	124	10	5.1.3.1	Ribulose-phosphate 3-epimerase, chloroplast precursor (EC 5.1.3.1) (Pentose-5-phosphate 3-epimerase) (PPE) (RPE) (R5P3E) (Fragment) - Solanum tuberosum (Potato)
1lbf	10.4	2.6	126	8	4.1.1.48	Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS) [Gene: trpC or SSO0895] - Sulfolobus solfataricus
1tqj	10.4	2.7	125	10	5.1.3.1	Ribulose-phosphate 3-epimerase (EC 5.1.3.1) (Pentose-5-phosphate 3-epimerase) (PPE) (R5P3E) [Gene: rpe or sll0807] - Synechocystis sp. (strain PCC 6803)
1lbl	10.2	2.6	125	9	4.1.1.48	Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS) [Gene: trpC or SSO0895] - Sulfolobus solfataricus
1igs	10.2	2.7	126	9	4.1.1.48	Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS) [Gene: trpC or SSO0895] - Sulfolobus solfataricus
1a53	10.1	2.6	125	9	4.1.1.48	Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS) [Gene: trpC or SSO0895] - Sulfolobus solfataricus

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1juk	10.1	2.7	126	8	4.1.1.48	Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS) [Gene: trpC or SSO0895] - Sulfolobus solfataricus
1vc4	9.9	2.5	124	5	not found	Crystal Structure of Indole-3-Glycerol Phosphate Synthase (TrpC) from Thermus Thermophilus At 1.8 A Resolution
1jcm	9.7	2.7	124	8	5.3.1.24	Tryptophan biosynthesis protein trpCF [Includes: Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS); N-(5'-phospho-ribosyl)anthranilate isomerase (EC 5.3.1.24) (PRAI)] [Gene: trpC or b1262 or JW1254] - Escherichia coli (strain K12)
1wq5	9.5	3.1	121	13	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or b1260 or JW1252] - Escherichia coli (strain K12)
1lbm	9.3	2.7	118	10	5.3.1.24	N-(5'-phosphoribosyl)anthranilate isomerase (EC 5.3.1.24) (PRAI) [Gene: trpF or TM_0139] - Thermotoga maritima
1fq0	9.2	2.6	121	13	4.1.2.14	KHG/KDPG aldolase [Includes: 4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (2-keto-4-hydroxyglutarate aldolase) (KHG-aldolase); 2-dehydro-3-deoxy-phosphogluconate aldolase (EC 4.1.2.14) (Phospho-2-dehydro-3-deoxygluconate aldolase) (Phospho-2-keto-3-deoxygluconate aldolase) (2-keto-3-deoxy-6-phosphogluconate aldolase) (KDPG-aldolase)] [Gene: eda or hga or kdgA or b1850 or JW1839] - Escherichia coli (strain K12)
1g67	9.1	2.4	113	11	2.5.1.3	Thiamine-phosphate pyrophosphorylase (EC 2.5.1.3) (TMP pyrophosphorylase) (TMP-PPase) (Thiamine-phosphate synthase) [Gene: thiE or thiC or ywbK or BSU38290 or ipa-26d] - Bacillus subtilis
1eua	9.1	2.6	119	13	4.1.2.14	KHG/KDPG aldolase [Includes: 4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (2-keto-4-hydroxyglutarate aldolase) (KHG-aldolase); 2-dehydro-3-deoxy-phosphogluconate aldolase (EC 4.1.2.14) (Phospho-2-dehydro-3-deoxygluconate aldolase) (Phospho-2-keto-3-deoxygluconate aldolase) (2-keto-3-deoxy-6-phosphogluconate aldolase) (KDPG-aldolase)] [Gene: eda or hga or kdgA or b1850 or JW1839] - Escherichia coli (strain K12)
1geq	9.1	2.8	123	11	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or PF1705] - Pyrococcus furiosus
1i4n	9.1	2.9	123	7	4.1.1.48	Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS) [Gene: trpC or TM_0140] - Thermotoga maritima
1g4t	9	2.3	112	10	2.5.1.3	Thiamine-phosphate pyrophosphorylase (EC 2.5.1.3) (TMP pyrophosphorylase) (TMP-PPase) (Thiamine-phosphate synthase) [Gene: thiE or thiC or ywbK or BSU38290 or ipa-26d] - Bacillus subtilis
1g6c	9	2.3	112	10	2.5.1.3	Thiamine-phosphate pyrophosphorylase (EC 2.5.1.3) (TMP pyrophosphorylase) (TMP-PPase) (Thiamine-phosphate synthase) [Gene: thiE or thiC or ywbK or BSU38290 or ipa-26d] - Bacillus subtilis
1g69	9	2.4	114	10	2.5.1.3	Thiamine-phosphate pyrophosphorylase (EC 2.5.1.3) (TMP pyrophosphorylase) (TMP-PPase) (Thiamine-phosphate synthase) [Gene: thiE or thiC or ywbK or BSU38290 or ipa-26d] - Bacillus subtilis
1g4s	9	2.4	113	10	2.5.1.3	Thiamine-phosphate pyrophosphorylase (EC 2.5.1.3) (TMP pyrophosphorylase) (TMP-PPase) (Thiamine-phosphate synthase) [Gene: thiE or thiC or ywbK or BSU38290 or ipa-26d] - Bacillus subtilis

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1fwr	9	2.5	119	14	4.1.3.16	KHG/KDPG aldolase [Includes: 4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (2-keto-4-hydroxyglutarate aldolase) (KHG-aldolase); 2-dehydro-3-deoxy-phosphogluconate aldolase (EC 4.1.2.14) (Phospho-2-dehydro-3-deoxygluconate aldolase) (Phospho-2-keto-3-deoxygluconate aldolase) (2-keto-3-deoxy-6-phosphogluconate aldolase) (KDPG-aldolase)] [Gene: eda or hga or kdgA or b1850 or JW1839] - Escherichia coli (strain K12)
1eun	9	2.5	119	13	4.1.3.16	KHG/KDPG aldolase [Includes: 4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (2-keto-4-hydroxyglutarate aldolase) (KHG-aldolase); 2-dehydro-3-deoxy-phosphogluconate aldolase (EC 4.1.2.14) (Phospho-2-dehydro-3-deoxygluconate aldolase) (Phospho-2-keto-3-deoxygluconate aldolase) (2-keto-3-deoxy-6-phosphogluconate aldolase) (KDPG-aldolase)] [Gene: eda or hga or kdgA or b1850 or JW1839] - Escherichia coli (strain K12)
1xxx	8.9	2.4	120	7	4.2.1.52	Dihydrodipicolinate synthase (EC 4.2.1.52) (DHDPS) [Gene: dapA or Rv2753c or MT2823 or MTV002.18c] - Mycobacterium tuberculosis
1fwr	8.8	2.5	118	14	4.1.2.14	KHG/KDPG aldolase [Includes: 4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (2-keto-4-hydroxyglutarate aldolase) (KHG-aldolase); 2-dehydro-3-deoxy-phosphogluconate aldolase (EC 4.1.2.14) (Phospho-2-dehydro-3-deoxygluconate aldolase) (Phospho-2-keto-3-deoxygluconate aldolase) (2-keto-3-deoxy-6-phosphogluconate aldolase) (KDPG-aldolase)] [Gene: eda or hga or kdgA or b1850 or JW1839] - Escherichia coli (strain K12)
1kfc	8.8	3	123	12	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1h1y	8.7	2.4	110	10	5.1.3.1	Ribulose-phosphate 3-epimerase, cytoplasmic isoform (EC 5.1.3.1) (Ribulose-5-phosphate-epimerase) (Cyt-RPEase) (RPEcyt) (Pentose-5-phosphate 3-epimerase) (PPE) [Gene: Os09g0505700 or LOC_Os09g32810] - Oryza sativa subsp. japonica (Rice)
1xxx	8.7	2.4	120	7	4.2.1.52	Dihydrodipicolinate synthase (EC 4.2.1.52) (DHDPS) [Gene: dapA or Rv2753c or MT2823 or MTV002.18c] - Mycobacterium tuberculosis
1vhc	8.7	2.6	121	13	4.1.3.16	Putative KHG/KDPG aldolase [Includes: 4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (2-keto-4-hydroxyglutarate aldolase) (KHG-aldolase); 2-dehydro-3-deoxy-phosphogluconate aldolase (EC 4.1.2.14) (Phospho-2-dehydro-3-deoxygluconate aldolase) (Phospho-2-keto-3-deoxygluconate aldolase) (2-keto-3-deoxy-6-phosphogluconate aldolase) (KDPG-aldolase)] [Gene: eda or HI0047] - Haemophilus influenzae
1xl9	8.7	2.6	121	17	not found	Crystal Structure of Dihydrodipicolinate Synthase DapA-2 (BA3935) from Bacillus Anthracis.
1xl9	8.6	2.5	121	17	not found	Crystal Structure of Dihydrodipicolinate Synthase DapA-2 (BA3935) from Bacillus Anthracis.
1h1y	8.5	2.3	107	10	5.1.3.1	Ribulose-phosphate 3-epimerase, cytoplasmic isoform (EC 5.1.3.1) (Ribulose-5-phosphate-epimerase) (Cyt-RPEase) (RPEcyt) (Pentose-5-phosphate 3-epimerase) (PPE) [Gene: Os09g0505700 or LOC_Os09g32810] - Oryza sativa subsp. japonica (Rice)
1n7k	8.5	2.5	102	20	4.1.2.4	Probable deoxyribose-phosphate aldolase (EC 4.1.2.4) (Phosphodeoxyriboaldolase) (Deoxyriboaldolase) (DERA) [Gene: deoC or APE_2437.1] - Aeropyrum pernix

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1s38	8.5	3	124	14	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1q63	8.5	3.1	123	14	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1p0d	8.5	3.1	124	14	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1h1z	8.4	2.4	108	10	5.1.3.1	Ribulose-phosphate 3-epimerase, cytoplasmic isoform (EC 5.1.3.1) (Ribulose-5-phosphate-epimerase) (Cyt-RPEase) (RPEcyt) (Pentose-5-phosphate 3-epimerase) (PPE) [Gene: Os09g0505700 or LOC_Os09g32810] - Oryza sativa subsp. japonica (Rice)
1q66	8.4	3	123	14	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1a5b	8.4	3	121	12	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1bks	8.4	3	120	13	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1p0b	8.4	3.1	124	14	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1q2r	8.4	3.2	126	14	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1f73	8.3	2.5	118	13	4.1.3.3	N-acetylneuraminate lyase (EC 4.1.3.3) (N-acetylneuraminic acid aldolase) (N-acetylneuraminate pyruvate-lyase) (Sialic acid lyase) (Sialate lyase) (Sialic acid aldolase) [Gene: nanA or HI0142] - Haemophilus influenzae
1q4w	8.3	3.1	124	14	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1n2v	8.3	3.1	124	14	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1km1	8.2	2.7	115	10	4.1.1.23	Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23) (OMP decarboxylase) (OMPDCase) (OMPdecase) [Gene: pyrF or MTH_129] - Methanobacterium thermoautotrophicum
1km2	8.2	2.7	115	10	4.1.1.23	Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23) (OMP decarboxylase) (OMPDCase) (OMPdecase) [Gene: pyrF or MTH_129] - Methanobacterium thermoautotrophicum
1km3	8.2	2.7	115	10	4.1.1.23	Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23) (OMP decarboxylase) (OMPDCase) (OMPdecase) [Gene: pyrF or MTH_129] - Methanobacterium thermoautotrophicum
1dvj	8.2	2.7	115	10	4.1.1.23	Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23) (OMP decarboxylase) (OMPDCase) (OMPdecase) [Gene: pyrF or MTH_129] - Methanobacterium thermoautotrophicum
2a6n	8.1	2.5	121	12	4.2.1.52	Dihydrodipicolinate synthase (EC 4.2.1.52) (DHDPS) [Gene: dapA or b2478 or JW2463] - Escherichia coli (strain K12)
1km0	8.1	2.6	114	10	4.1.1.23	Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23) (OMP decarboxylase) (OMPDCase) (OMPdecase) [Gene: pyrF or MTH_129] - Methanobacterium thermoautotrophicum
1gvf	8.1	2.7	119	12	4.1.2.-	Tagatose-1,6-bisphosphate aldolase kbaY (EC 4.1.2.-) (TBPA) [Gene: kbaY or agaY or kba or yraC or b3137 or JW3106] - Escherichia coli (strain K12)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1aj0	8.1	3	109	11	2.5.1.15	Dihydropteroate synthase (EC 2.5.1.15) (DHPS) (Dihydropteroate pyrophosphorylase) [Gene: folP or dhpS or b3177 or JW3144] - Escherichia coli (strain K12)
1y5v	8	3	118	15	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1jcl	7.9	2.6	105	14	4.1.2.4	Deoxyribose-phosphate aldolase (EC 4.1.2.4) (Phosphodeoxyriboaldolase) (Deoxyriboaldolase) (DERA) [Gene: deoC or dra or thyR or b4381 or JW4344] - Escherichia coli (strain K12)
1so4	7.9	2.6	103	8	4.1.1.85	3-keto-L-gulonate-6-phosphate decarboxylase ulaD (EC 4.1.1.85) (3-dehydro-L-gulonate-6-phosphate decarboxylase) (KGPDC) (L-ascorbate utilization protein D) [Gene: ulaD or sgaH or yjfV or b4196 or JW4154] - Escherichia coli (strain K12)
1ad4	7.9	2.7	110	14	2.5.1.15	Dihydropteroate synthase (EC 2.5.1.15) (Dihydropteroate pyrophosphorylase) (DHPS) [Gene: folP] - Staphylococcus aureus
1jcj	7.9	2.7	107	14	4.1.2.4	Deoxyribose-phosphate aldolase (EC 4.1.2.4) (Phosphodeoxyriboaldolase) (Deoxyriboaldolase) (DERA) [Gene: deoC or dra or thyR or b4381 or JW4344] - Escherichia coli (strain K12)
1so3	7.8	2.6	103	10	4.1.1.85	3-keto-L-gulonate-6-phosphate decarboxylase ulaD (EC 4.1.1.85) (3-dehydro-L-gulonate-6-phosphate decarboxylase) (KGPDC) (L-ascorbate utilization protein D) [Gene: ulaD or sgaH or yjfV or b4196 or JW4154] - Escherichia coli (strain K12)
1ad1	7.8	2.6	110	14	2.5.1.15	Dihydropteroate synthase (EC 2.5.1.15) (Dihydropteroate pyrophosphorylase) (DHPS) [Gene: folP] - Staphylococcus aureus
1xl9	7.8	2.6	119	16	not found	Crystal Structure of Dihydrodipicolinate Synthase DapA-2 (BA3935) from Bacillus Anthracis.
1q6q	7.7	2.7	105	10	4.1.1.85	3-keto-L-gulonate-6-phosphate decarboxylase ulaD (EC 4.1.1.85) (3-dehydro-L-gulonate-6-phosphate decarboxylase) (KGPDC) (L-ascorbate utilization protein D) [Gene: ulaD or sgaH or yjfV or b4196 or JW4154] - Escherichia coli (strain K12)
1wx0	7.7	2.9	112	15	not found	Crystal structure of transaldolase from Thermus thermophilus HB8
1f7b	7.6	2.2	97	14	4.1.3.3	N-acetylneuraminate lyase (EC 4.1.3.3) (N-acetylneuraminic acid aldolase) (N-acetylneuraminate pyruvate-lyase) (Sialic acid lyase) (Sialate lyase) (Sialic acid aldolase) [Gene: nanA or HI0142] - Haemophilus influenzae
1f6p	7.6	2.2	97	14	4.1.3.3	N-acetylneuraminate lyase (EC 4.1.3.3) (N-acetylneuraminic acid aldolase) (N-acetylneuraminate pyruvate-lyase) (Sialic acid lyase) (Sialate lyase) (Sialic acid aldolase) [Gene: nanA or HI0142] - Haemophilus influenzae
1f6k	7.6	2.2	97	14	4.1.3.3	N-acetylneuraminate lyase (EC 4.1.3.3) (N-acetylneuraminic acid aldolase) (N-acetylneuraminate pyruvate-lyase) (Sialic acid lyase) (Sialate lyase) (Sialic acid aldolase) [Gene: nanA or HI0142] - Haemophilus influenzae
1f5z	7.6	2.2	97	14	4.1.3.3	N-acetylneuraminate lyase (EC 4.1.3.3) (N-acetylneuraminic acid aldolase) (N-acetylneuraminate pyruvate-lyase) (Sialic acid lyase) (Sialate lyase) (Sialic acid aldolase) [Gene: nanA or HI0142] - Haemophilus influenzae

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1xbx	7.6	2.6	102	10	4.1.1.85	3-keto-L-gulonate-6-phosphate decarboxylase ulaD (EC 4.1.1.85) (3-dehydro-L-gulonate-6-phosphate decarboxylase) (KGPDC) (L-ascorbate utilization protein D) [Gene: ulaD or sgaH or yjfV or b4196 or JW4154] - Escherichia coli (strain K12)
1twz	7.6	2.7	106	17	not found	Dihydropteroate Synthetase, With Bound Substrate Analogue PtP, From Bacillus anthracis
1ktn	7.6	2.7	105	14	4.1.2.4	Deoxyribose-phosphate aldolase (EC 4.1.2.4) (Phosphodeoxyriboaldolase) (Deoxyriboaldolase) (DERA) [Gene: deoC or dra or thyR or b4381 or JW4344] - Escherichia coli (strain K12)
1fdy	7.5	2.1	97	10	4.1.3.3	N-acetylneuraminate lyase (EC 4.1.3.3) (N-acetylneuraminic acid aldolase) (N-acetylneuraminate pyruvate-lyase) (Sialic acid lyase) (Sialate lyase) (Sialic acid aldolase) (NALase) [Gene: nanA or npl or b3225 or JW3194] - Escherichia coli (strain K12)
1hl2	7.5	2.2	97	10	4.1.3.3	N-acetylneuraminate lyase (EC 4.1.3.3) (N-acetylneuraminic acid aldolase) (N-acetylneuraminate pyruvate-lyase) (Sialic acid lyase) (Sialate lyase) (Sialic acid aldolase) (NALase) [Gene: nanA or npl or b3225 or JW3194] - Escherichia coli (strain K12)
1fdz	7.5	2.2	97	10	4.1.3.3	N-acetylneuraminate lyase (EC 4.1.3.3) (N-acetylneuraminic acid aldolase) (N-acetylneuraminate pyruvate-lyase) (Sialic acid lyase) (Sialate lyase) (Sialic acid aldolase) (NALase) [Gene: nanA or npl or b3225 or JW3194] - Escherichia coli (strain K12)
1q26	7.5	2.7	107	17	not found	not found
1tws	7.5	2.7	107	17	not found	Dihydropteroate Synthetase From Bacillus anthracis
1k4g	7.4	3.2	114	14	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1kw1	7.3	2.7	101	9	4.1.1.85	3-keto-L-gulonate-6-phosphate decarboxylase ulaD (EC 4.1.1.85) (3-dehydro-L-gulonate-6-phosphate decarboxylase) (KGPDC) (L-ascorbate utilization protein D) [Gene: ulaD or sgaH or yjfV or b4196 or JW4154] - Escherichia coli (strain K12)
1n8w	7.1	2.4	98	13	2.3.3.9	Malate synthase G (EC 2.3.3.9) [Gene: glcB or Rv1837c or MT1885 or MTCY1A11.06] - Mycobacterium tuberculosis
1eye	7.1	2.8	103	15	2.5.1.15	Dihydropteroate synthase 1 (EC 2.5.1.15) (DHPS 1) (Dihydropteroate pyrophosphorylase 1) [Gene: folP1 or Rv3608c or MT3712 or MTCY07H7B.14] - Mycobacterium tuberculosis
1b57	7	2.6	98	8	4.1.2.13	Fructose-bisphosphate aldolase class 2 (EC 4.1.2.13) (Fructose-bisphosphate aldolase class II) (FBP aldolase) [Gene: fbaA or fba or fda or b2925 or JW2892] - Escherichia coli (strain K12)
1tx2	7	2.8	100	17	not found	Dihydropteroate Synthetase, With Bound Inhibitor MANIC, From Bacillus anthracis
1rvg	6.8	2	96	14	not found	crystal strcuture of class II fructose-bisphosphate aldolase from Thermus aquaticus in complex with Y
1rv8	6.8	2	95	14	not found	Class II fructose-1,6-bisphosphate aldolase from Thermus aquaticus in complex with cobalt
1bkh	6.8	3	103	10	not found	MUCONATE LACTONIZING ENZYME FROM PSEUDOMONAS PUTIDA

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1g4e	6.7	2	86	9	2.5.1.3	Thiamine-phosphate pyrophosphorylase (EC 2.5.1.3) (TMP pyrophosphorylase) (TMP-PPase) (Thiamine-phosphate synthase) [Gene: thiE or thiC or ywbK or BSU38290 or ipa-26d] - Bacillus subtilis
1rv8	6.7	2	96	14	not found	Class II fructose-1,6-bisphosphate aldolase from Thermus aquaticus in complex with cobalt
1rv8	6.7	2	95	14	not found	Class II fructose-1,6-bisphosphate aldolase from Thermus aquaticus in complex with cobalt
1g4p	6.6	2.1	86	9	2.5.1.3	Thiamine-phosphate pyrophosphorylase (EC 2.5.1.3) (TMP pyrophosphorylase) (TMP-PPase) (Thiamine-phosphate synthase) [Gene: thiE or thiC or ywbK or BSU38290 or ipa-26d] - Bacillus subtilis
1dhp	6.6	2.5	98	9	4.2.1.52	Dihydrodipicolinate synthase (EC 4.2.1.52) (DHDPS) [Gene: dapA or b2478 or JW2463] - Escherichia coli (strain K12)
1mmf	6.6	3.2	116	11	not found	Crystal structure of substrate free form of glycerol dehydratase
1muc	6.5	2.6	99	10	5.5.1.1	Muconate cycloisomerase 1 (EC 5.5.1.1) (Muconate cycloisomerase I) (Cis,cis-muconate lactonizing enzyme I) (MLE) [Gene: catB] - Pseudomonas putida
1n8i	6.5	2.6	95	13	2.3.3.9	Malate synthase G (EC 2.3.3.9) [Gene: glcB or Rv1837c or MT1885 or MTCY1A11.06] - Mycobacterium tuberculosis
1qpq	6.5	2.8	95	12	2.4.2.19	Nicotinate-nucleotide pyrophosphorylase [carboxylating] (EC 2.4.2.19) (Quinolinate phosphoribosyltransferase [decarboxylating]) (QAPRTase) [Gene: nadC or Rv1596 or MT1632 or MTCY336.08c] - Mycobacterium tuberculosis
1iwb	6.3	3.2	112	9	not found	Crystal structure of diol dehydratase
1i2o	6.2	3	106	14	2.2.1.2	Transaldolase B (EC 2.2.1.2) [Gene: talB or yaaK or b0008 or JW0007] - Escherichia coli (strain K12)
1i2q	6.2	3	106	14	2.2.1.2	Transaldolase B (EC 2.2.1.2) [Gene: talB or yaaK or b0008 or JW0007] - Escherichia coli (strain K12)
1it8	5.9	2.9	95	14	2.4.2.-	7-cyano-7-deazaguanine tRNA-ribosyltransferase (EC 2.4.2.-) (Archaeal tRNA-guanine transglycosylase) [Gene: tgtA or PH1116] - Pyrococcus horikoshii
1j2b	5.8	2.9	96	14	2.4.2.-	7-cyano-7-deazaguanine tRNA-ribosyltransferase (EC 2.4.2.-) (Archaeal tRNA-guanine transglycosylase) [Gene: tgtA or PH1116] - Pyrococcus horikoshii
1onr	5.8	2.9	106	14	2.2.1.2	Transaldolase B (EC 2.2.1.2) [Gene: talB or yaaK or b0008 or JW0007] - Escherichia coli (strain K12)
1iq8	5.7	3.1	97	13	2.4.2.-	7-cyano-7-deazaguanine tRNA-ribosyltransferase (EC 2.4.2.-) (Archaeal tRNA-guanine transglycosylase) [Gene: tgtA or PH1116] - Pyrococcus horikoshii
1ec7	5.6	2.7	99	9	4.2.1.40	Glucarate dehydratase (EC 4.2.1.40) (GDH) (GlucD) [Gene: gudD or ygcX or b2787 or JW2758] - Escherichia coli (strain K12)
1ucw	5.6	3.1	107	12	2.2.1.2	Transaldolase B (EC 2.2.1.2) [Gene: talB or yaaK or b0008 or JW0007] - Escherichia coli (strain K12)
1zpt	5.6	3.2	93	10	1.5.1.20	5,10-methylenetetrahydrofolate reductase (EC 1.5.1.20) [Gene: metF or b3941 or JW3913] - Escherichia coli (strain K12)
1o4u	5.4	2.6	88	20	not found	Crystal structure of type II quinolic acid phosphoribosyltransferase (TM1645) from Thermotoga maritima at 2.50 A resolution
1ecq	5.4	3.2	103	9	4.2.1.40	Glucarate dehydratase (EC 4.2.1.40) (GDH) (GlucD) [Gene: gudD or ygcX or b2787 or JW2758] - Escherichia coli (strain K12)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1b5t	5.3	2.8	98	11	1.5.1.20	5,10-methylenetetrahydrofolate reductase (EC 1.5.1.20) [Gene: metF or b3941 or JW3913] - Escherichia coli (strain K12)
1pii	5.2	2.6	70	4	4.1.1.48	Tryptophan biosynthesis protein trpCF [Includes: Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS); N-(5'-phospho-ribosyl)anthranilate isomerase (EC 5.3.1.24) (PRAI)] [Gene: trpC or b1262 or JW1254] - Escherichia coli (strain K12)
1jdf	5.2	3.4	104	11	4.2.1.40	Glucarate dehydratase (EC 4.2.1.40) (GDH) (GlucD) [Gene: gudD or ygcX or b2787 or JW2758] - Escherichia coli (strain K12)
1ec9	5	3.1	99	8	4.2.1.40	Glucarate dehydratase (EC 4.2.1.40) (GDH) (GlucD) [Gene: gudD or ygcX or b2787 or JW2758] - Escherichia coli (strain K12)
1p7t	4.9	2.3	79	9	2.3.3.9	Malate synthase G (EC 2.3.3.9) (MSG) [Gene: glcB or glc or b2976 or JW2943] - Escherichia coli (strain K12)
1a80	4.9	3.1	86	7	1.1.1.274	2,5-diketo-D-gluconic acid reductase A (EC 1.1.1.274) (2,5-DKG reductase A) (2,5-DKGR A) (25DKGR-A) (AKR5C) [Gene: dkgA] - Corynebacterium sp. (strain ATCC 31090)
1igw	4.8	2.5	79	6	4.1.3.1	Isocitrate lyase (EC 4.1.3.1) (Isocitrase) (Isocitratase) (ICL) [Gene: aceA or icl or b4015 or JW3975] - Escherichia coli (strain K12)
1tuf	4.8	2.7	84	11	4.1.1.20	Diaminopimelate decarboxylase (EC 4.1.1.20) (DAP decarboxylase) [Gene: lysA or MJ1097] - Methanocaldococcus jannaschii (Methanococcus jannaschii)
1v93	4.8	2.9	98	11	not found	5,10-Methylenetetrahydrofolate Reductase from Thermus thermophilus HB8
1dos	4.8	3.3	78	5	4.1.2.13	Fructose-bisphosphate aldolase class 2 (EC 4.1.2.13) (Fructose-bisphosphate aldolase class II) (FBP aldolase) [Gene: fbaA or fba or fda or b2925 or JW2892] - Escherichia coli (strain K12)
1a50	4.7	2.9	102	7	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1bgg	4.6	2.6	85	6	3.2.1.21	Beta-glucosidase A (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) (Amygdalase) (BGA) [Gene: bglA] - Paenibacillus polymyxa (Bacillus polymyxa)
1bga	4.6	2.7	86	6	3.2.1.21	Beta-glucosidase A (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) (Amygdalase) (BGA) [Gene: bglA] - Paenibacillus polymyxa (Bacillus polymyxa)
1fba	4.6	3.3	97	6	4.1.2.13	Fructose-bisphosphate aldolase (EC 4.1.2.13) [Gene: Ald or CG6058] - Drosophila melanogaster (Fruit fly)
1kko	4.5	2.7	80	6	not found	CRYSTAL STRUCTURE OF CITROBACTER AMALONATICUS METHYLASPARTATE AMMONIA LYASE
1zrq	4.5	2.8	97	11	1.5.1.20	5,10-methylenetetrahydrofolate reductase (EC 1.5.1.20) [Gene: metF or b3941 or JW3913] - Escherichia coli (strain K12)
1vpx	4.5	2.8	83	8	2.2.1.2	Transaldolase (EC 2.2.1.2) [Gene: tal or TM_0295] - Thermotoga maritima
1f8m	4.4	2.7	83	8	4.1.3.1	Isocitrate lyase (EC 4.1.3.1) (Isocitrase) (Isocitratase) (ICL) [Gene: icl or Rv0467 or MT0483 or MTV038.11] - Mycobacterium tuberculosis
1gon	4.4	2.7	86	7	not found	B-GLUCOSIDASE FROM STREPTOMYCES SP
1a5c	4.4	3.1	91	10	4.1.2.13	Fructose-bisphosphate aldolase (EC 4.1.2.13) (41 kDa antigen) - Plasmodium falciparum

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1bqc	4.3	2.9	90	6	not found	BETA-MANNANASE FROM THERMOMONOSPORA FUSCA
1tv5	4.3	3	86	9	1.3.3.1	Dihydroorotate dehydrogenase homolog, mitochondrial precursor (EC 1.3.3.1) (Dihydroorotate oxidase) (DHOdehase) [Gene: PFF0160c] - Plasmodium falciparum (isolate 3D7)
1c3f	4.3	3.8	92	12	3.2.1.96	Endo-beta-N-acetylglucosaminidase H precursor (EC 3.2.1.96) (Mannosyl-glycoprotein endo-beta-N-acetyl-glucosaminidase H) (DI-N-acetylchitobiosyl beta-N-acetylglucosaminidase H) (Endoglycosidase H) (Endo H) - Streptomyces plicatus
1eex	4.2	2.8	101	4	not found	CRYSTAL STRUCTURE OF THE DIOL DEHYDRATASE-ADENINYLPENTYLCOBALAMIN COMPLEX FROM KLEBSIELLA OXYTOCA
1wkf	4.2	2.8	72	13	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1wkd	4.2	2.9	73	11	2.4.2.29	Queuine tRNA-ribosyltransferase (EC 2.4.2.29) (tRNA-guanine transglycosylase) (Guanine insertion enzyme) [Gene: tgt or ZMO0363] - Zymomonas mobilis
1ewd	4.2	3	82	9	4.1.2.13	Fructose-bisphosphate aldolase A (EC 4.1.2.13) (Muscle-type aldolase) [Gene: ALDOA] - Oryctolagus cuniculus (Rabbit)
1b3x	4.2	3.2	82	11	3.2.1.8	Endo-1,4-beta-xylanase (EC 3.2.1.8) (Xylanase) (1,4-beta-D-xylan xylanohydrolase) - Penicillium simplicissimum
1b3z	4.2	3.3	83	11	3.2.1.8	Endo-1,4-beta-xylanase (EC 3.2.1.8) (Xylanase) (1,4-beta-D-xylan xylanohydrolase) - Penicillium simplicissimum
1b30	4.2	3.3	83	12	3.2.1.8	Endo-1,4-beta-xylanase (EC 3.2.1.8) (Xylanase) (1,4-beta-D-xylan xylanohydrolase) - Penicillium simplicissimum
1hfb	4.2	3.4	82	6	2.5.1.54	Phospho-2-dehydro-3-deoxyheptonate aldolase, tyrosine-inhibited (EC 2.5.1.54) (Phospho-2-keto-3-deoxyheptonate aldolase) (DAHP synthetase) (3-deoxy-D-arabino-heptulosonate 7-phosphate synthase) [Gene: ARO4 or YBR249C or YBR1701] - Saccharomyces cerevisiae (Baker's yeast)
1bgl	4.2	3.5	93	10	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1geq	4.1	1.9	66	11	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or PF1705] - Pyrococcus furiosus
1l6w	4.1	2.8	83	6	4.1.2.-	Fructose-6-phosphate aldolase 1 (EC 4.1.2.-) [Gene: fsaA or fsa or mipB or ybiZ or b0825 or JW5109] - Escherichia coli (strain K12)
1c7t	4.1	2.8	86	6	3.2.1.52	Chitobiase precursor (EC 3.2.1.52) (N-acetyl-beta-glucosaminidase) (Beta-N-acetylhexosaminidase) [Gene: chb] - Serratia marcescens
1gyn	4.1	3.1	97	7	4.1.2.13	Fructose-bisphosphate aldolase class 2 (EC 4.1.2.13) (Fructose-bisphosphate aldolase class II) (FBP aldolase) [Gene: fbaA or fba or fda or b2925 or JW2892] - Escherichia coli (strain K12)
1xdm	4.1	3.1	93	9	4.1.2.13	Fructose-bisphosphate aldolase B (EC 4.1.2.13) (Liver-type aldolase) [Gene: ALDOB or ALDB] - Homo sapiens (Human)
1b3v	4.1	3.2	82	11	3.2.1.8	Endo-1,4-beta-xylanase (EC 3.2.1.8) (Xylanase) (1,4-beta-D-xylan xylanohydrolase) - Penicillium simplicissimum
1d7k	4.1	3.2	103	5	not found	Ornithine decarboxylase (EC 4.1.1.17) (ODC) [Gene: ODC1] - Homo sapiens (Human)
1af7	31.2	0			2.1.1.80	Chemotaxis protein methyltransferase (EC 2.1.1.80) [Gene: cheR or STM1918] - Salmonella typhimurium

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1bc5	25.2	0.3	136	100	2.1.1.80	Chemotaxis protein methyltransferase (EC 2.1.1.80) [Gene: cheR or STM1918] - Salmonella typhimurium
1dus	10.9	2.3	111	15	0.0.0.0	Protein MJ0882 [Gene: MJ0882] - Methanocaldococcus jannaschii (Methanococcus jannaschii)
2avn	10.3	2.2	107	24	not found	Crystal structure of Ubiquinone/menaquinone biosynthesis methyltransferase-related protein (tm1389) from THERMOTOGA MARITIMA at 2.35 A resolution
1qan	9.2	2.6	109	12	2.1.1.48	rRNA adenine N-6-methyltransferase (EC 2.1.1.48) (Macrolide-lincosamide-streptogramin B resistance protein) (Erythromycin resistance protein) [Gene: ermC'] - Bacillus subtilis
1qaq	9.1	2.7	111	15	2.1.1.48	rRNA adenine N-6-methyltransferase (EC 2.1.1.48) (Macrolide-lincosamide-streptogramin B resistance protein) (Erythromycin resistance protein) [Gene: ermC'] - Bacillus subtilis
2erc	8.7	2.3	100	15	2.1.1.48	rRNA adenine N-6-methyltransferase (EC 2.1.1.48) (Macrolide-lincosamide-streptogramin B resistance protein) (Erythromycin resistance protein) [Gene: ermC'] - Bacillus subtilis
1d2c	8.6	1.9	83	20	2.1.1.20	Glycine N-methyltransferase (EC 2.1.1.20) (Folate-binding protein) [Gene: Gnmt] - Rattus norvegicus (Rat)
1xva	8.6	2.2	85	21	2.1.1.20	Glycine N-methyltransferase (EC 2.1.1.20) (Folate-binding protein) [Gene: Gnmt] - Rattus norvegicus (Rat)
1r74	8.5	2.4	85	20	2.1.1.20	Glycine N-methyltransferase (EC 2.1.1.20) [Gene: GNMT] - Homo sapiens (Human)
1d2g	8.5	2.7	85	21	2.1.1.20	Glycine N-methyltransferase (EC 2.1.1.20) (Folate-binding protein) [Gene: Gnmt] - Rattus norvegicus (Rat)
1bhj	8.3	3.3	83	22	2.1.1.20	Glycine N-methyltransferase (EC 2.1.1.20) (Folate-binding protein) [Gene: Gnmt] - Rattus norvegicus (Rat)
1vbf	7.9	2.9	106	16	not found	Crystal structure of protein L-isoaspartate O-methyltransferase homologue from Sulfolobus tokodaii
1qyr	7.8	2.7	104	4	2.1.1.-	Dimethyladenosine transferase (EC 2.1.1.-) (S-adenosylmethionine-6-N', N'-adenosyl(rRNA) dimethyltransferase) (16S rRNA dimethylase) (High level kasugamycin resistance protein ksgA) (Kasugamycin dimethyltransferase) [Gene: ksgA or rsmA or b0051 or JW0050] - Escherichia coli (strain K12)
1aqj	7.2	2.6	101	21	2.1.1.72	Modification methylase TaqI (EC 2.1.1.72) (Adenine-specific methyltransferase TaqI) (M.TaqI) [Gene: taqIM] - Thermus aquaticus
1aqi	7.1	2.7	102	21	2.1.1.72	Modification methylase TaqI (EC 2.1.1.72) (Adenine-specific methyltransferase TaqI) (M.TaqI) [Gene: taqIM] - Thermus aquaticus
1g38	6.7	2.6	101	21	2.1.1.72	Modification methylase TaqI (EC 2.1.1.72) (Adenine-specific methyltransferase TaqI) (M.TaqI) [Gene: taqIM] - Thermus aquaticus
2b3t	6.5	2.4	85	14	2.1.1.-	Protein methyltransferase hemK (EC 2.1.1.-) (Protein-glutamine N-methyltransferase hemK) (Protein-(glutamine-N(5)) MTase hemK) (M.EcoKHemKP) [Gene: hemK or prmC or b1212 or JW1203] - Escherichia coli (strain K12)
2fyt	6	2.9	93	17	2.1.1.-	Protein arginine N-methyltransferase 3 (EC 2.1.1.-) (Heterogeneous nuclear ribonucleoprotein methyltransferase-like protein 3) [Gene: PRMT3 or HRMT1L3] - Homo sapiens (Human)
1xds	5.8	1.3	63	19	not found	Crystal structure of Aclacinomycin-10-hydroxylase (RdmB) in complex with S-adenosyl-L-methionine (SAM) and 11-deoxy-beta-rhodomycin (DbrA)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1t43	5.8	3.1	79	14	2.1.1.-	Protein methyltransferase hemK (EC 2.1.1.-) (Protein-glutamine N-methyltransferase hemK) (Protein-(glutamine-N(5)) MTase hemK) (M.EcoKHemKP) [Gene: hemK or prmC or b1212 or JW1203] - Escherichia coli (strain K12)
1bw0	5.5	3.2	92	13	2.6.1.5	Tyrosine aminotransferase (EC 2.6.1.5) (L-tyrosine:2-oxoglutarate aminotransferase) (TAT) - Trypanosoma cruzi
1xdu	5.4	1.4	69	17	not found	Crystal structure of Aclacinomycin-10-hydroxylase (RdmB) in complex with Sinefungin (SFG)
1orh	4.9	3	84	21	2.1.1.-	Protein arginine N-methyltransferase 1 (EC 2.1.1.-) [Gene: Prmt1 or Hrmt1l2] - Rattus norvegicus (Rat)
1g6q	4.7	3.1	85	19	2.1.1.-	HNRNP arginine N-methyltransferase (EC 2.1.1.-) (Protein ODP1) [Gene: HMT1 or ODP1 or RMT1 or YBR034C or YBR0320] - Saccharomyces cerevisiae (Baker's yeast)
1or8	4.6	3	83	22	2.1.1.-	Protein arginine N-methyltransferase 1 (EC 2.1.1.-) [Gene: Prmt1 or Hrmt1l2] - Rattus norvegicus (Rat)
1wzn	4.5	1.5	62	27	not found	Crystal Structure of the SAM-dependent methyltransferase from Pyrococcus horikoshii OT3
1cl2	4.2	2.9	88	8	4.4.1.8	Cystathionine beta-lyase (EC 4.4.1.8) (CBL) (Beta-cystathionase) (Cysteine lyase) [Gene: metC or b3008 or JW2975] - Escherichia coli (strain K12)
1aod	28	0			4.6.1.13	1-phosphatidylinositol phosphodiesterase precursor (EC 4.6.1.13) (Phosphatidylinositol diacylglycerol-lyase) (Phosphatidylinositol-specific phospholipase C) (PI-PLC) [Gene: plcA or pic or lmo0201] - Listeria monocytogenes
2ptd	15.1	1.5	98	37	4.6.1.13	1-phosphatidylinositol phosphodiesterase precursor (EC 4.6.1.13) (Phosphatidylinositol diacylglycerol-lyase) (Phosphatidylinositol-specific phospholipase C) (PI-PLC) - Bacillus cereus
1ptd	14.5	1.6	98	36	4.6.1.13	1-phosphatidylinositol phosphodiesterase precursor (EC 4.6.1.13) (Phosphatidylinositol diacylglycerol-lyase) (Phosphatidylinositol-specific phospholipase C) (PI-PLC) - Bacillus cereus
7ptd	14.2	1.4	94	36	4.6.1.13	1-phosphatidylinositol phosphodiesterase precursor (EC 4.6.1.13) (Phosphatidylinositol diacylglycerol-lyase) (Phosphatidylinositol-specific phospholipase C) (PI-PLC) - Bacillus cereus
1djw	9.3	2.6	100	22	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)
1djy	9.3	2.7	101	23	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)
2isd	9.1	2.7	100	22	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)
1qas	9	2.2	92	25	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)
1qas	9	2.4	96	23	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1djz	8.8	2.4	96	25	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)
1qat	8.7	2.4	94	24	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)
1djh	8.6	2.4	93	25	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)
1dji	8.6	2.5	95	23	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)
1djg	8.6	2.5	94	23	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)
1djx	8.4	2.7	100	22	3.1.4.11	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III) [Gene: Plcd1] - Rattus norvegicus (Rat)
1az9	37	0			not found	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
2bws	35.6	0.1	178	99	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
1wl6	35.5	0.1	177	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
1wlr	35.4	0.2	177	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
2bn7	35.2	0.2	178	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
2bwy	35	0.2	177	99	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
2bww	34.1	0.2	175	99	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
1m35	33.8	0.2	177	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
1w2m	33.7	0.2	178	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1n51	32.8	0.2	174	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
1w7v	29.3	0.2	155	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
1wbq	29	0.2	154	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
2bwu	29	0.2	155	99	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
1wn1	25.7	1.3	165	35	not found	Crystal Structure of Dipeptiase from Pyrococcus Horikoshii OT3
1pv9	21.1	1.1	144	40	3.4.13.9	Xaa-Pro dipeptidase (EC 3.4.13.9) (X-Pro dipeptidase) (Proline dipeptidase) (Prolidase) (Imidodipeptidase) [Gene: pepQ or PF1343] - Pyrococcus furiosus
1wkm	20.2	1.9	157	22	3.4.11.18	Methionine aminopeptidase (EC 3.4.11.18) (MAP) (Peptidase M) [Gene: map or PF0541] - Pyrococcus furiosus
1o0x	19.6	1.9	156	29	not found	Crystal structure of Methionine aminopeptidase (TM1478) from Thermotoga maritima at 1.90 A resolution
1c22	17.5	1.8	143	27	3.4.11.18	Methionine aminopeptidase (EC 3.4.11.18) (MAP) (Peptidase M) [Gene: map or b0168 or JW0163] - Escherichia coli (strain K12)
1c27	17.4	1.8	143	27	3.4.11.18	Methionine aminopeptidase (EC 3.4.11.18) (MAP) (Peptidase M) [Gene: map or b0168 or JW0163] - Escherichia coli (strain K12)
1c24	17.4	1.8	143	27	3.4.11.18	Methionine aminopeptidase (EC 3.4.11.18) (MAP) (Peptidase M) [Gene: map or b0168 or JW0163] - Escherichia coli (strain K12)
1qxz	17.3	2	147	24	3.4.11.18	Methionine aminopeptidase (EC 3.4.11.18) (MAP) (Peptidase M) [Gene: map or SAV1888] - Staphylococcus aureus (strain Mu50 / ATCC 700699)
1xnz	17	1.8	144	26	3.4.11.18	Methionine aminopeptidase (EC 3.4.11.18) (MAP) (Peptidase M) [Gene: map or b0168 or JW0163] - Escherichia coli (strain K12)
1yvm	16	1.9	139	26	3.4.11.18	Methionine aminopeptidase (EC 3.4.11.18) (MAP) (Peptidase M) [Gene: map or b0168 or JW0163] - Escherichia coli (strain K12)
1r58	15.3	2.1	141	17	3.4.11.18	Methionine aminopeptidase 2 (EC 3.4.11.18) (MetAP 2) (MAP 2) (Peptidase M 2) (Initiation factor 2-associated 67 kDa glycoprotein) (p67) (p67eIF2) [Gene: METAP2 or MNPEP or P67EIF2] - Homo sapiens (Human)
1xgn	15.1	2	134	20	3.4.11.18	Methionine aminopeptidase (EC 3.4.11.18) (MAP) (Peptidase M) [Gene: map or PF0541] - Pyrococcus furiosus
1r5g	15.1	2.1	139	17	3.4.11.18	Methionine aminopeptidase 2 (EC 3.4.11.18) (MetAP 2) (MAP 2) (Peptidase M 2) (Initiation factor 2-associated 67 kDa glycoprotein) (p67) (p67eIF2) [Gene: METAP2 or MNPEP or P67EIF2] - Homo sapiens (Human)
1qzy	15	2.1	139	17	3.4.11.18	Methionine aminopeptidase 2 (EC 3.4.11.18) (MetAP 2) (MAP 2) (Peptidase M 2) (Initiation factor 2-associated 67 kDa glycoprotein) (p67) (p67eIF2) [Gene: METAP2 or MNPEP or P67EIF2] - Homo sapiens (Human)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1bn5	14.9	2.1	135	17	3.4.11.18	Methionine aminopeptidase 2 (EC 3.4.11.18) (MetAP 2) (MAP 2) (Peptidase M 2) (Initiation factor 2-associated 67 kDa glycoprotein) (p67) (p67eIF2) [Gene: METAP2 or MNPEP or P67EIF2] - Homo sapiens (Human)
1b6a	14.9	2.2	135	17	3.4.11.18	Methionine aminopeptidase 2 (EC 3.4.11.18) (MetAP 2) (MAP 2) (Peptidase M 2) (Initiation factor 2-associated 67 kDa glycoprotein) (p67) (p67eIF2) [Gene: METAP2 or MNPEP or P67EIF2] - Homo sapiens (Human)
1yj3	14.7	1.8	135	27	3.4.11.18	Methionine aminopeptidase (EC 3.4.11.18) (MAP) (Peptidase M) [Gene: map or mapB or Rv2861c or MT2929 or MTV003.07c] - Mycobacterium tuberculosis
1b59	14.7	2.1	135	17	3.4.11.18	Methionine aminopeptidase 2 (EC 3.4.11.18) (MetAP 2) (MAP 2) (Peptidase M 2) (Initiation factor 2-associated 67 kDa glycoprotein) (p67) (p67eIF2) [Gene: METAP2 or MNPEP or P67EIF2] - Homo sapiens (Human)
1chm	14	1.5	126	22	3.5.3.3	Creatinase (EC 3.5.3.3) (Creatine amidinohydrolase) - Pseudomonas putida
1kp0	13.6	1.5	122	24	not found	The Crystal Structure Analysis of Creatine Amidinohydrolase from Actinobacillus
1qxw	11.5	2.2	127	24	3.4.11.18	Methionine aminopeptidase (EC 3.4.11.18) (MAP) (Peptidase M) [Gene: map or SAV1888] - Staphylococcus aureus (strain Mu50 / ATCC 700699)
2bha	11.4	0.2	76	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
2bhc	11.4	0.2	76	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
2bh3	11.3	0.2	76	100	3.4.11.9	Xaa-Pro aminopeptidase (EC 3.4.11.9) (X-Pro aminopeptidase) (Aminopeptidase P II) (APP-II) (Aminoacylproline aminopeptidase) [Gene: pepP or b2908 or JW2876] - Escherichia coli (strain K12)
1b3q	30.1	0			2.7.13.3	Chemotaxis protein cheA (EC 2.7.13.3) [Gene: cheA or TM_0702] - Thermotoga maritima
1i5c	23.9	1.2	137	99	2.7.13.3	Chemotaxis protein cheA (EC 2.7.13.3) [Gene: cheA or TM_0702] - Thermotoga maritima
1i58	23.6	1.5	142	94	2.7.13.3	Chemotaxis protein cheA (EC 2.7.13.3) [Gene: cheA or TM_0702] - Thermotoga maritima
1i5a	22.2	0.9	124	98	2.7.13.3	Chemotaxis protein cheA (EC 2.7.13.3) [Gene: cheA or TM_0702] - Thermotoga maritima
1i5b	22.1	0.9	123	98	2.7.13.3	Chemotaxis protein cheA (EC 2.7.13.3) [Gene: cheA or TM_0702] - Thermotoga maritima
1i59	21.8	1.2	124	98	2.7.13.3	Chemotaxis protein cheA (EC 2.7.13.3) [Gene: cheA or TM_0702] - Thermotoga maritima
1i5d	20.9	1.7	128	98	2.7.13.3	Chemotaxis protein cheA (EC 2.7.13.3) [Gene: cheA or TM_0702] - Thermotoga maritima
1y8o	13.5	2	115	17	2.7.11.2	Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex, mitochondrial precursor (EC 2.3.1.12) (Pyruvate dehydrogenase complex E2 subunit) (PDCE2) (E2) (Dihydrolipoamide S-acetyltransferase component of pyruvate dehydrogenase complex) (PDC-E2) (70 kDa mitochondrial autoantigen of primary biliary cirrhosis) (PBC) (M2 antigen complex 70 kDa subunit) [Gene: DLAT or DLTA] - Homo sapiens (Human)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
2btz	12.4	1.8	107	22	2.7.11.2	[Pyruvate dehydrogenase [lipoamide]] kinase isozyme 2, mitochondrial precursor (EC 2.7.11.2) (Pyruvate dehydrogenase kinase isoform 2) [Gene: PDK2] - Homo sapiens (Human)
2bu8	12.3	1.8	107	21	2.7.11.2	[Pyruvate dehydrogenase [lipoamide]] kinase isozyme 2, mitochondrial precursor (EC 2.7.11.2) (Pyruvate dehydrogenase kinase isoform 2) [Gene: PDK2] - Homo sapiens (Human)
2bu6	12.3	1.8	107	22	2.7.11.2	[Pyruvate dehydrogenase [lipoamide]] kinase isozyme 2, mitochondrial precursor (EC 2.7.11.2) (Pyruvate dehydrogenase kinase isoform 2) [Gene: PDK2] - Homo sapiens (Human)
1jm6	11.5	2.1	104	21	not found	[Pyruvate dehydrogenase [lipoamide]] kinase isozyme 2, mitochondrial precursor (EC 2.7.11.2) (Pyruvate dehydrogenase kinase isoform 2) (PDK P45) [Gene: Pdk2] - Rattus norvegicus (Rat)
1gkz	10.1	2.2	109	22	2.7.11.4	[3-methyl-2-oxobutanoate dehydrogenase [lipoamide]] kinase, mitochondrial precursor (EC 2.7.11.4) (Branched-chain alpha-ketoacid dehydrogenase kinase) (BCKDHKIN) (BCKD-kinase) [Gene: Bckdk] - Rattus norvegicus (Rat)
1bgl	31.6	0			3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1f49	28.9	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1gho	28.8	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jz1	28.7	0.2	142	100	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jz0	28.6	0.2	142	100	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1dp0	28.5	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1hn1	28.4	0.3	143	100	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jyw	28.4	1.1	144	98	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jz3	28.4	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jz7	28.4	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jz8	28.3	1.1	144	98	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jz5	28	1.1	143	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1px4	27.7	0.3	142	100	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1f4a	27.5	0.2	143	100	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1f4h	27.4	0.3	143	100	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1bgl	27.3	0.2	138	100	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1px3	27.3	0.3	141	100	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1bgm	27.3	0.3	143	100	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1hn1	27.3	0.5	141	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1hn1	27.2	0.5	140	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jyz	27.2	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jyy	27.2	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1hn1	27	0.3	138	100	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jyx	26.9	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jz6	26.8	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jz2	26.8	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jz4	26.7	1.1	144	99	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jyv	26.7	1.1	144	98	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1jyn	26.7	1.1	144	98	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacZ or b0344 or JW0335] - Escherichia coli (strain K12)
1yq2	16.9	2	121	51	not found	beta-galactosidase from Arthrobacter sp. C2-2 (isoenzyme C2-2-1)
1yq2	16.9	2.1	121	51	not found	beta-galactosidase from Arthrobacter sp. C2-2 (isoenzyme C2-2-1)
1yq2	15.9	2	118	53	not found	beta-galactosidase from Arthrobacter sp. C2-2 (isoenzyme C2-2-1)
1yq2	15.9	2.1	121	51	not found	beta-galactosidase from Arthrobacter sp. C2-2 (isoenzyme C2-2-1)
1yq2	15.8	2.1	121	51	not found	beta-galactosidase from Arthrobacter sp. C2-2 (isoenzyme C2-2-1)
1yq2	15.4	2.1	121	51	not found	beta-galactosidase from Arthrobacter sp. C2-2 (isoenzyme C2-2-1)
1h1n	8.2	2.4	100	18	not found	ATOMIC RESOLUTION STRUCTURE OF THE MAJOR ENDOGLUCANASE FROM THERMOASCUS AURANTIACUS
1gzj	8.2	2.5	100	19	not found	STRUCTURE OF THERMOASCUS AURANTIACUS FAMILY 5 ENDOGLUCANASE
1bhg	8	2.2	90	31	3.2.1.31	Beta-glucuronidase precursor (EC 3.2.1.31) (Beta-G1) [Gene: GUSB] - Homo sapiens (Human)
1gzj	8	2.4	101	19	not found	STRUCTURE OF THERMOASCUS AURANTIACUS FAMILY 5 ENDOGLUCANASE
1hf5	7.2	2.8	100	15	not found	Endoglucanase 5A (EC 3.2.1.4) (Endo-1,4-beta-glucanase 5A) (Alkaline cellulase) [Gene: cel5A] - Bacillus agaradhaerens (Bacillus agaradherans)
1egz	7.1	2.7	100	14	3.2.1.4	Endoglucanase Z precursor (EC 3.2.1.4) (Endo-1,4-beta-glucanase Z) (Cellulase Z) (EGZ) [Gene: celZ or cel5] - Dickeya dadantii (strain 3937) (Erwinia chrysanthemi (strain 3937))
2c0h	7.1	2.8	104	12	3.2.1.78	Mannan endo-1,4-beta-mannosidase precursor (EC 3.2.1.78) (Beta-mannanase) (Endo-beta-1,4-mannanase) (ManA) - Mytilus edulis (Blue mussel)
1fhl	7.1	2.9	106	17	3.2.1.89	Arabinogalactan endo-1,4-beta-galactosidase precursor (EC 3.2.1.89) (Endo-1,4-beta-galactanase) (Galactanase) [Gene: gal1] - Aspergillus aculeatus

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1lf1	7	2.9	100	16	not found	Crystal Structure of Cel5 from Alkalophilic Bacillus sp.
1gon	6.8	2.7	97	19	not found	B-GLUCOSIDASE FROM STREPTOMYCES SP
1h5v	6.8	2.7	100	15	3.2.1.4	Endoglucanase 5A (EC 3.2.1.4) (Endo-1,4-beta-glucanase 5A) (Alkaline cellulase) [Gene: cel5A] - Bacillus agaradhaerens (Bacillus agaradherans)
1e5j	6.8	3.1	103	15	3.2.1.4	Endoglucanase B (EC 3.2.1.4) (Endo-1,4-beta-glucanase B) (Cellulase B) [Gene: celB] - Bacillus sp. (strain N-4 / JCM 9156)
1edg	6.7	2.5	95	19	3.2.1.4	Endoglucanase A precursor (EC 3.2.1.4) (Endo-1,4-beta-glucanase A) (Cellulase A) (EGCCA) [Gene: celCCA] - Clostridium cellulolyticum
1gnx	6.7	2.7	97	19	not found	B-GLUCOSIDASE FROM STREPTOMYCES SP
1hf7	6.6	2.8	100	15	not found	Endoglucanase 5A (EC 3.2.1.4) (Endo-1,4-beta-glucanase 5A) (Alkaline cellulase) [Gene: cel5A] - Bacillus agaradhaerens (Bacillus agaradherans)
1h11	6.6	2.8	100	15	3.2.1.4	Endoglucanase 5A (EC 3.2.1.4) (Endo-1,4-beta-glucanase 5A) (Alkaline cellulase) [Gene: cel5A] - Bacillus agaradhaerens (Bacillus agaradherans)
1hf6	6.6	2.8	100	15	3.2.1.4	Endoglucanase 5A (EC 3.2.1.4) (Endo-1,4-beta-glucanase 5A) (Alkaline cellulase) [Gene: cel5A] - Bacillus agaradhaerens (Bacillus agaradherans)
1k6a	6.6	2.9	97	16	3.2.1.8	Endo-1,4-beta-xylanase precursor (EC 3.2.1.8) (Xylanase) (1,4-beta-D-xylan xylanohydrolase) (TAXI) [Gene: XYNA] - Thermoascus aurantiacus
1g01	6.6	3	100	16	3.2.1.4	Endoglucanase precursor (EC 3.2.1.4) (Endo-1,4-beta-glucanase) (Alkaline cellulase) - Bacillus sp. (strain KSM-635)
1g0c	6.6	3	100	16	3.2.1.4	Endoglucanase precursor (EC 3.2.1.4) (Endo-1,4-beta-glucanase) (Alkaline cellulase) - Bacillus sp. (strain KSM-635)
1goo	6.6	3.2	101	17	3.2.1.8	Endo-1,4-beta-xylanase precursor (EC 3.2.1.8) (Xylanase) (1,4-beta-D-xylan xylanohydrolase) (TAXI) [Gene: XYNA] - Thermoascus aurantiacus
1np2	6.5	2.7	96	19	not found	Crystal structure of thermostable beta-glycosidase from thermophilic eubacterium Thermus nonproteolyticus HG102
1qox	6.5	2.7	97	14	3.2.1.21	Beta-glucosidase (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) (Amygdalase) [Gene: bglA] - Bacillus circulans
1.00E+73	6.5	2.7	97	10	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1dwa	6.5	2.7	97	10	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1dwf	6.5	2.7	97	10	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1e4m	6.5	2.7	97	10	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1dwg	6.5	2.7	97	10	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1dwi	6.5	2.7	97	10	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1c0d	6.5	2.8	94	16	not found	Endoglucanase E1 precursor (EC 3.2.1.4) (Endo-1,4-beta-glucanase E1) (Cellulase E1) (Endocellulase E1) [Gene: Acel_0614] - Acidothermus cellulolyticus (strain ATCC 43068 / 11B)
1gow	6.5	2.9	100	14	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacS or SSO3019] - Sulfolobus solfataricus
1gor	6.5	3.1	100	17	3.2.1.8	Endo-1,4-beta-xylanase precursor (EC 3.2.1.8) (Xylanase) (1,4-beta-D-xylan xylanohydrolase) (TAXI) [Gene: XYNA] - Thermoascus aurantiacus
1.00E+56	6.4	2.6	99	14	3.2.1.21	Beta-glucosidase, chloroplast precursor (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) [Gene: GLU1] - Zea mays (Maize)
1w9d	6.4	2.7	97	10	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1.00E+72	6.4	2.9	98	10	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1it0	6.4	3.1	96	14	not found	Crystal structure of xylanase from Streptomyces olivaceoviridis E-86 complexed with lactose
1e0v	6.4	3.1	96	14	3.2.1.8	Endo-1,4-beta-xylanase A precursor (EC 3.2.1.8) (Xylanase A) (1,4-beta-D-xylan xylanohydrolase A) [Gene: xlnA] - Streptomyces lividans
1.00E+55	6.3	2.7	98	13	3.2.1.21	Beta-glucosidase, chloroplast precursor (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) [Gene: GLU1] - Zea mays (Maize)
1e4n	6.3	2.7	99	12	3.2.1.21	Beta-glucosidase, chloroplast precursor (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) [Gene: GLU1] - Zea mays (Maize)
1cbg	6.3	2.7	96	11	3.2.1.21	Cyanogenic beta-glucosidase precursor (EC 3.2.1.21) (Linamarase) (Fragment) [Gene: LI] - Trifolium repens (Creeping white clover)
1e4l	6.3	2.7	98	13	3.2.1.21	Beta-glucosidase, chloroplast precursor (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) [Gene: GLU1] - Zea mays (Maize)
1e6q	6.3	2.8	97	11	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1e1e	6.3	2.8	101	14	3.2.1.21	Beta-glucosidase, chloroplast precursor (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) [Gene: GLU1] - Zea mays (Maize)
1e0x	6.3	2.9	94	14	3.2.1.8	Endo-1,4-beta-xylanase A precursor (EC 3.2.1.8) (Xylanase A) (1,4-beta-D-xylan xylanohydrolase A) [Gene: xlnA] - Streptomyces lividans
1i1w	6.3	2.9	97	16	3.2.1.8	Endo-1,4-beta-xylanase precursor (EC 3.2.1.8) (Xylanase) (1,4-beta-D-xylan xylanohydrolase) (TAXI) [Gene: XYNA] - Thermoascus aurantiacus
1isz	6.3	3	96	15	not found	Crystal structure of xylanase from Streptomyces olivaceoviridis E-86 complexed with galactose
1hjs	6.3	3	104	16	3.2.1.89	Arabinogalactan endo-1,4-beta-galactosidase (EC 3.2.1.89) (Endo-1,4-beta-galactanase) (Galactanase) - Thielavia heterothallica (Myceliophthora thermophila)
1isy	6.3	3.1	97	13	not found	Crystal structure of xylanase from Streptomyces olivaceoviridis E-86 complexed with glucose

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1isx	6.3	3.1	97	13	not found	Crystal structure of xylanase from Streptomyces olivaceoviridis E-86 complexed with xylotriose
1isw	6.3	3.1	97	13	not found	Crystal structure of xylanase from Streptomyces olivaceoviridis E-86 complexed with xylobiose
1isv	6.3	3.2	98	13	not found	Crystal structure of xylanase from Streptomyces olivaceoviridis E-86 complexed with xylose
1.00E+70	6.2	2.8	94	11	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1hxj	6.2	2.8	99	13	3.2.1.21	Beta-glucosidase, chloroplast precursor (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) [Gene: GLU1] - Zea mays (Maize)
1dwh	6.1	2.8	94	11	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
2cbu	6.1	2.8	96	20	3.2.1.21	Beta-glucosidase A (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) [Gene: bglA] - Thermotoga maritima
1h49	6.1	2.8	101	13	3.2.1.21	Beta-glucosidase, chloroplast precursor (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) [Gene: GLU1] - Zea mays (Maize)
1e6s	6.1	2.8	93	11	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1e1f	6.1	2.9	102	14	3.2.1.21	Beta-glucosidase, chloroplast precursor (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) [Gene: GLU1] - Zea mays (Maize)
1gow	6.1	2.9	100	13	3.2.1.23	Beta-galactosidase (EC 3.2.1.23) (Lactase) [Gene: lacS or SSO3019] - Sulfolobus solfataricus
1e6x	6	2.8	92	11	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1w9b	6	2.8	92	11	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1.00E+71	6	2.8	92	11	3.2.1.147	Myrosinase MA1 (EC 3.2.1.147) (Sinigrinase) (Thioglucosidase) - Sinapis alba (White mustard) (Brassica hirta)
1tax	6	2.9	97	16	not found	Endo-1,4-beta-xylanase precursor (EC 3.2.1.8) (Xylanase) (1,4-beta-D-xylan xylanohydrolase) (TAXI) [Gene: XYNA] - Thermoascus aurantiacus
1ug6	5.9	2.7	96	20	not found	Structure of beta-glucosidase at atomic resolution from thermus thermophilus HB8
1exp	5.8	3	94	18	3.2.1.91	Exoglucanase/xylanase precursor [Includes: Exoglucanase (EC 3.2.1.91) (Exocellobiohydrolase) (1,4-beta-cellobiohydrolase) (Beta-1,4-glycanase CEX); Endo-1,4-beta-xylanase B (EC 3.2.1.8) (Xylanase B)] [Gene: cex or xynB] - Cellulomonas fimi
1fh9	5.8	3.1	95	18	not found	CRYSTAL STRUCTURE OF THE XYLANASE CEX WITH XYLOBIOSE-DERIVED LACTAM OXIME INHIBITOR
1fh7	5.8	3.4	97	18	not found	CRYSTAL STRUCTURE OF THE XYLANASE CEX WITH XYLOBIOSE-DERIVED INHIBITOR DEOXYNOJIRIMYCIN
1v6w	5.7	2.9	93	14	not found	Crystal Structure Of Xylanase From Streptomyces Olivaceoviridis E-86 Complexed With 2(2)-4-O-methyl-alpha-D-glucuronosyl-xylobiose
1hiz	5.5	3.3	97	10	3.2.1.8	Endo-1,4-beta-xylanase precursor (EC 3.2.1.8) (Xylanase) (1,4-beta-D-xylan xylanohydrolase) - Bacillus stearothermophilus (Geobacillus stearothermophilus)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
3a3h	5.2	2.7	82	17	3.2.1.4	Endoglucanase 5A (EC 3.2.1.4) (Endo-1,4-beta-glucanase 5A) (Alkaline cellulase) [Gene: cel5A] - Bacillus agaradhaerens (Bacillus agaradherans)
1e4i	4.9	2.5	80	15	3.2.1.21	Beta-glucosidase A (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) (Amygdalase) (BGA) [Gene: bglA] - Paenibacillus polymyxa (Bacillus polymyxa)
2a3h	4.9	2.6	75	19	3.2.1.4	Endoglucanase 5A (EC 3.2.1.4) (Endo-1,4-beta-glucanase 5A) (Alkaline cellulase) [Gene: cel5A] - Bacillus agaradhaerens (Bacillus agaradherans)
1px8	4.8	2.8	92	15	3.2.1.37	Beta-xylosidase (EC 3.2.1.37) (1,4-beta-D-xylan xylohydrolase) (Xylan 1,4-beta-xylosidase) [Gene: xynB] - Thermoanaerobacter saccharolyticum
1fob	4.7	3.1	85	16	3.2.1.89	Arabinogalactan endo-1,4-beta-galactosidase precursor (EC 3.2.1.89) (Endo-1,4-beta-galactanase) (Galactanase) [Gene: gal1] - Aspergillus aculeatus
1uhv	4.6	3.1	96	14	3.2.1.37	Beta-xylosidase (EC 3.2.1.37) (1,4-beta-D-xylan xylohydrolase) (Xylan 1,4-beta-xylosidase) [Gene: xynB] - Thermoanaerobacter saccharolyticum
1uah	4.4	2.5	82	23	not found	not found
1fcv	4.2	3	88	14	3.2.1.35	Hyaluronoglucosaminidase precursor (EC 3.2.1.35) (Hyaluronidase) (Hya) (Allergen Api m 2) (Api m II) - Apis mellifera (Honeybee)
1clx	4.1	3.3	87	17	3.2.1.8	Endo-1,4-beta-xylanase A precursor (EC 3.2.1.8) (Xylanase A) (1,4-beta-D-xylan xylanohydrolase A) (XYLA) [Gene: xynA] - Pseudomonas fluorescens
1bif	32.6	0			2.7.1.105	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4 (6PF-2-K/Fru-2,6-P2ASE testis-type isozyme) [Includes: 6-phosphofructo-2-kinase (EC 2.7.1.105); Fructose-2,6-bisphosphatase (EC 3.1.3.46)] [Gene: Pfkfb4] - Rattus norvegicus (Rat)
1k6m	30.1	0.4	147	86	3.1.3.46	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1 (6PF-2-K/Fru-2,6-P2ASE liver isozyme) [Includes: 6-phosphofructo-2-kinase (EC 2.7.1.105); Fructose-2,6-bisphosphatase (EC 3.1.3.46)] [Gene: PFKFB1 or F6PK or PFRX] - Homo sapiens (Human)
1c81	27.3	0.6	141	84	2.7.1.105	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1 (6PF-2-K/Fru-2,6-P2ASE liver isozyme) [Includes: 6-phosphofructo-2-kinase (EC 2.7.1.105); Fructose-2,6-bisphosphatase (EC 3.1.3.46)] [Gene: Pfkfb1] - Rattus norvegicus (Rat)
1c7z	26.3	0.5	136	85	2.7.1.105	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1 (6PF-2-K/Fru-2,6-P2ASE liver isozyme) [Includes: 6-phosphofructo-2-kinase (EC 2.7.1.105); Fructose-2,6-bisphosphatase (EC 3.1.3.46)] [Gene: Pfkfb1] - Rattus norvegicus (Rat)
1c80	25.5	0.8	134	84	2.7.1.105	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1 (6PF-2-K/Fru-2,6-P2ASE liver isozyme) [Includes: 6-phosphofructo-2-kinase (EC 2.7.1.105); Fructose-2,6-bisphosphatase (EC 3.1.3.46)] [Gene: Pfkfb1] - Rattus norvegicus (Rat)
1fbt	25.2	0.5	132	86	2.7.1.105	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1 (6PF-2-K/Fru-2,6-P2ASE liver isozyme) [Includes: 6-phosphofructo-2-kinase (EC 2.7.1.105); Fructose-2,6-bisphosphatase (EC 3.1.3.46)] [Gene: Pfkfb1] - Rattus norvegicus (Rat)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
2bif	24.8	0.2	126	99	3.1.3.46	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4 (6PF-2-K/Fru-2,6-P2ASE testis-type isozyme) [Includes: 6-phosphofructo-2-kinase (EC 2.7.1.105); Fructose-2,6-bisphosphatase (EC 3.1.3.46)] [Gene: Pfkfb4] - Rattus norvegicus (Rat)
1tip	24.8	0.5	136	85	2.7.1.105	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1 (6PF-2-K/Fru-2,6-P2ASE liver isozyme) [Includes: 6-phosphofructo-2-kinase (EC 2.7.1.105); Fructose-2,6-bisphosphatase (EC 3.1.3.46)] [Gene: Pfkfb1] - Rattus norvegicus (Rat)
1.00E+58	19.2	1.7	135	28	5.4.2.1	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase (EC 5.4.2.1) (Phosphoglyceromutase) (PGAM) (BPG-dependent PGAM) (dPGM) [Gene: gpmA or gpm or pgm or pgmA or b0755 or JW0738] - Escherichia coli (strain K12)
1ebb	19.2	1.8	137	28	not found	BACILLUS STEAROTHERMOPHILUS YHFR
1qhf	19	2.1	141	24	5.4.2.1	Phosphoglycerate mutase 1 (EC 5.4.2.1) (Phosphoglyceromutase 1) (PGAM 1) (MPGM 1) (BPG-dependent PGAM 1) [Gene: GPM1 or GPM or YKL152C or YKL607] - Saccharomyces cerevisiae (Baker's yeast)
1h2e	18.9	1.9	140	28	not found	BACILLUS STEAROTHERMOPHILUS PHOE (PREVIOUSLY KNOWN AS YHFR) IN COMPLEX WITH PHOSPHATE
1qhf	18.7	2.3	138	24	5.4.2.1	Phosphoglycerate mutase 1 (EC 5.4.2.1) (Phosphoglyceromutase 1) (PGAM 1) (MPGM 1) (BPG-dependent PGAM 1) [Gene: GPM1 or GPM or YKL152C or YKL607] - Saccharomyces cerevisiae (Baker's yeast)
1rii	18.6	1.6	131	27	5.4.2.1	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase (EC 5.4.2.1) (Phosphoglyceromutase) (PGAM) (BPG-dependent PGAM) (dPGM) [Gene: gpmA or gpm or gpm1 or pgm or Rv0489 or MT0508 or MTCY20G9.15] - Mycobacterium tuberculosis
1t8p	18.4	1.8	141	28	5.4.2.1	Bisphosphoglycerate mutase (EC 5.4.2.4) (2,3-bisphosphoglycerate mutase, erythrocyte) (2,3-bisphosphoglycerate synthase) (BPGM) (EC 5.4.2.1) (EC 3.1.3.13) (BPG-dependent PGAM) [Gene: BPGM] - Homo sapiens (Human)
1yjx	18.3	1.8	134	31	5.4.2.4	Phosphoglycerate mutase 1 (EC 5.4.2.1) (EC 5.4.2.4) (EC 3.1.3.13) (Phosphoglycerate mutase isozyme B) (PGAM-B) (BPG-dependent PGAM 1) [Gene: PGAM1 or PGAMA or CDABP0006] - Homo sapiens (Human)
1xq9	18.2	1.8	135	30	not found	Structure of Phosphoglycerate Mutase from Plasmodium falciparum at 2.6 Resolution
1xq9	17.5	1.7	135	30	not found	Structure of Phosphoglycerate Mutase from Plasmodium falciparum at 2.6 Resolution
4pgm	17.5	1.7	127	25	5.4.2.1	Phosphoglycerate mutase 1 (EC 5.4.2.1) (Phosphoglyceromutase 1) (PGAM 1) (MPGM 1) (BPG-dependent PGAM 1) [Gene: GPM1 or GPM or YKL152C or YKL607] - Saccharomyces cerevisiae (Baker's yeast)
1bq4	17.4	1.8	130	25	5.4.2.1	Phosphoglycerate mutase 1 (EC 5.4.2.1) (Phosphoglyceromutase 1) (PGAM 1) (MPGM 1) (BPG-dependent PGAM 1) [Gene: GPM1 or GPM or YKL152C or YKL607] - Saccharomyces cerevisiae (Baker's yeast)
1v37	17.4	1.9	126	25	not found	Crystal structure of phosphoglycerate mutase from Thermus thermophilus HB8
1bq3	17.1	1.8	128	25	5.4.2.1	Phosphoglycerate mutase 1 (EC 5.4.2.1) (Phosphoglyceromutase 1) (PGAM 1) (MPGM 1) (BPG-dependent PGAM 1) [Gene: GPM1 or GPM or YKL152C or YKL607] - Saccharomyces cerevisiae (Baker's yeast)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1yfk	16.9	1.8	129	30	3.1.3.13	Phosphoglycerate mutase 1 (EC 5.4.2.1) (EC 5.4.2.4) (EC 3.1.3.13) (Phosphoglycerate mutase isozyme B) (PGAM-B) (BPG-dependent PGAM 1) [Gene: PGAM1 or PGAMA or CDABP0006] - Homo sapiens (Human)
1fzt	16.7	2.1	139	24	5.4.2.1	Phosphoglycerate mutase (EC 5.4.2.1) (Phosphoglyceromutase) (PGAM) (MPGM) (BPG-dependent PGAM) [Gene: gpm1 or SPAC26F1.06] - Schizosaccharomyces pombe (Fission yeast)
1v7q	16	1.9	125	25	not found	Crystal structure of phosphoglycerate mutase from Thermus thermophilus HB8
5pgm	15.7	1.8	113	29	5.4.2.1	Phosphoglycerate mutase 1 (EC 5.4.2.1) (Phosphoglyceromutase 1) (PGAM 1) (MPGM 1) (BPG-dependent PGAM 1) [Gene: GPM1 or GPM or YKL152C or YKL607] - Saccharomyces cerevisiae (Baker's yeast)
1cvi	9.4	3	95	18	3.1.3.2	Prostatic acid phosphatase precursor (EC 3.1.3.2) [Gene: ACPP] - Homo sapiens (Human)
1qfx	7.5	2.6	92	16	3.1.3.8	3-phytase B precursor (EC 3.1.3.8) (Myo-inositol-hexaphosphate 3-phosphohydrolase B) (pH 2.5 optimum acid phosphatase) [Gene: phyB or aph] - Aspergillus awamori
1bmt	21.3	0			2.1.1.13	Methionine synthase (EC 2.1.1.13) (5-methyltetrahydrofolate--homocysteine methyltransferase) (Methionine synthase, vitamin-B12-dependent) (MS) [Gene: metH or b4019 or JW3979] - Escherichia coli (strain K12)
1k7y	16.3	0.4	73	99	2.1.1.13	Methionine synthase (EC 2.1.1.13) (5-methyltetrahydrofolate--homocysteine methyltransferase) (Methionine synthase, vitamin-B12-dependent) (MS) [Gene: metH or b4019 or JW3979] - Escherichia coli (strain K12)
1k98	15.5	0.5	73	99	2.1.1.13	Methionine synthase (EC 2.1.1.13) (5-methyltetrahydrofolate--homocysteine methyltransferase) (Methionine synthase, vitamin-B12-dependent) (MS) [Gene: metH or b4019 or JW3979] - Escherichia coli (strain K12)
1y80	14.4	0.7	72	36	not found	Structure of a corrinoid (factor IIIm)-binding protein from Moorella thermoacetica
1i9c	8.9	1.3	65	29	5.4.99.1	Methylaspartate mutase E chain (EC 5.4.99.1) (Glutamate mutase subunit epsilon) [Gene: glmE] - Clostridium cochlearium
1ccw	8.6	1.2	63	30	5.4.99.1	Methylaspartate mutase E chain (EC 5.4.99.1) (Glutamate mutase subunit epsilon) [Gene: glmE] - Clostridium cochlearium
1cb7	8.5	1.3	64	30	5.4.99.1	Methylaspartate mutase E chain (EC 5.4.99.1) (Glutamate mutase subunit epsilon) [Gene: glmE] - Clostridium cochlearium
1req	8.4	1.7	77	22	5.4.99.2	Methylmalonyl-CoA mutase small subunit (EC 5.4.99.2) (MCB-beta) [Gene: mutA] - Propionibacterium freudenreichii subsp. shermanii
2req	8.4	2	79	22	5.4.99.2	Methylmalonyl-CoA mutase small subunit (EC 5.4.99.2) (MCB-beta) [Gene: mutA] - Propionibacterium freudenreichii subsp. shermanii
1e1c	8.3	1.6	77	22	5.4.99.2	Methylmalonyl-CoA mutase small subunit (EC 5.4.99.2) (MCB-beta) [Gene: mutA] - Propionibacterium freudenreichii subsp. shermanii
4req	8.3	1.8	79	22	5.4.99.2	Methylmalonyl-CoA mutase small subunit (EC 5.4.99.2) (MCB-beta) [Gene: mutA] - Propionibacterium freudenreichii subsp. shermanii

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
5req	8.3	1.8	79	22	5.4.99.2	Methylmalonyl-CoA mutase small subunit (EC 5.4.99.2) (MCB-beta) [Gene: mutA] - Propionibacterium freudenreichii subsp. shermanii
6req	8.3	1.8	79	22	5.4.99.2	Methylmalonyl-CoA mutase small subunit (EC 5.4.99.2) (MCB-beta) [Gene: mutA] - Propionibacterium freudenreichii subsp. shermanii
7req	8.3	1.8	79	22	5.4.99.2	Methylmalonyl-CoA mutase small subunit (EC 5.4.99.2) (MCB-beta) [Gene: mutA] - Propionibacterium freudenreichii subsp. shermanii
1fmf	7.8	1.9	64	30	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	7.4	2	63	29	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	7.3	1.9	60	30	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	7.3	2.1	71	24	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	7.3	2.1	65	29	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	7	1.9	61	26	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	7	2.2	63	29	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	7	2.3	64	30	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	6.8	2.1	69	26	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	6.7	2	61	31	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	6.5	1.8	63	30	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	6.5	2.1	64	28	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	6.4	1.8	61	30	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	6.4	2.2	63	29	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	6.3	2	63	29	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	6.2	1.9	58	28	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1fmf	6.2	2	64	30	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	6.2	2.1	64	28	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	6.1	2.5	63	29	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	5.7	2	61	30	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	5.6	2.1	60	27	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	5.5	2.2	63	30	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1fmf	5.5	2.3	63	24	5.4.99.1	Methylaspartate mutase S chain (EC 5.4.99.1) (Glutamate mutase subunit sigma) [Gene: mamA or mutS] - Clostridium tetanomorphum
1j04	4.5	2.3	65	9	2.6.1.44	Serine--pyruvate aminotransferase (EC 2.6.1.51) (SPT) (Alanine--glyoxylate aminotransferase) (EC 2.6.1.44) (AGT) [Gene: AGXT or AGT1 or SPAT] - Homo sapiens (Human)
1j1z	4.5	2.5	60	17	6.3.4.5	Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase) [Gene: argG or TTHA0284] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1kh2	4.5	2.6	61	16	6.3.4.5	Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase) [Gene: argG or TTHA0284] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1j20	4.4	2.4	60	17	6.3.4.5	Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase) [Gene: argG or TTHA0284] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1kh2	4.4	2.5	61	16	6.3.4.5	Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase) [Gene: argG or TTHA0284] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1kh1	4.4	2.5	60	17	6.3.4.5	Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase) [Gene: argG or TTHA0284] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1nn4	4.4	2.7	68	18	5.3.1.6	Ribose-5-phosphate isomerase B (EC 5.3.1.6) (Phosphoriboisomerase B) [Gene: rpiB or yjcA or b4090 or JW4051] - Escherichia coli (strain K12)
1kh1	4.3	2.6	61	16	6.3.4.5	Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase) [Gene: argG or TTHA0284] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1tqx	4.2	2	52	15	not found	Crystal Structure of Pfal009167 A Putative D-Ribulose 5-Phosphate 3-Epimerase from P.falciparum
2ase	4.2	2.3	58	12	0.0.0.0	Cell division control protein 42 homolog precursor (G25K GTP-binding protein) [Gene: CDC42] - Homo sapiens (Human)
1kh2	4.2	2.6	60	15	6.3.4.5	Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase) [Gene: argG or TTHA0284] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1j21	4.2	2.6	60	15	6.3.4.5	Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase) [Gene: argG or TTHA0284] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
2afh	4.1	2.3	61	10	1.18.6.1	Nitrogenase molybdenum-iron protein alpha chain (EC 1.18.6.1) (Nitrogenase component I) (Dinitrogenase) [Gene: nifD] - Azotobacter vinelandii

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1kh1	4.1	2.5	59	15	6.3.4.5	Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase) [Gene: argG or TTHA0284] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1bxr	37.7	0			6.3.5.5	Carbamoyl-phosphate synthase small chain (EC 6.3.5.5) (Carbamoyl-phosphate synthetase glutamine chain) [Gene: carA or pyrA or b0032 or JW0030] - Escherichia coli (strain K12)
1c30	34.8	0.2	166	99	6.3.5.5	Carbamoyl-phosphate synthase small chain (EC 6.3.5.5) (Carbamoyl-phosphate synthetase glutamine chain) [Gene: carA or pyrA or b0032 or JW0030] - Escherichia coli (strain K12)
1c3o	34.6	0.2	167	99	6.3.5.5	Carbamoyl-phosphate synthase small chain (EC 6.3.5.5) (Carbamoyl-phosphate synthetase glutamine chain) [Gene: carA or pyrA or b0032 or JW0030] - Escherichia coli (strain K12)
1t36	30	0.2	147	100	6.3.5.5	Carbamoyl-phosphate synthase small chain (EC 6.3.5.5) (Carbamoyl-phosphate synthetase glutamine chain) [Gene: carA or pyrA or b0032 or JW0030] - Escherichia coli (strain K12)
1ce8	29.8	0.2	146	100	6.3.5.5	Carbamoyl-phosphate synthase small chain (EC 6.3.5.5) (Carbamoyl-phosphate synthetase glutamine chain) [Gene: carA or pyrA or b0032 or JW0030] - Escherichia coli (strain K12)
1jdb	29.4	0.2	146	100	6.3.5.5	Carbamoyl-phosphate synthase small chain (EC 6.3.5.5) (Carbamoyl-phosphate synthetase glutamine chain) [Gene: carA or pyrA or b0032 or JW0030] - Escherichia coli (strain K12)
1m6v	23.5	1.5	139	99	6.3.5.5	Carbamoyl-phosphate synthase small chain (EC 6.3.5.5) (Carbamoyl-phosphate synthetase glutamine chain) [Gene: carA or pyrA or b0032 or JW0030] - Escherichia coli (strain K12)
1cs0	22.8	0.2	128	100	6.3.5.5	Carbamoyl-phosphate synthase small chain (EC 6.3.5.5) (Carbamoyl-phosphate synthetase glutamine chain) [Gene: carA or pyrA or b0032 or JW0030] - Escherichia coli (strain K12)
1kee	22.6	0.2	128	100	6.3.5.5	Carbamoyl-phosphate synthase small chain (EC 6.3.5.5) (Carbamoyl-phosphate synthetase glutamine chain) [Gene: carA or pyrA or b0032 or JW0030] - Escherichia coli (strain K12)
1qdl	21.2	1.8	149	26	4.1.3.27	Anthranilate synthase component 1 (EC 4.1.3.27) (Anthranilate synthase component I) [Gene: trpE or SSO0893] - Sulfolobus solfataricus
2a9v	20	2.1	147	26	6.3.5.2	GMP synthase [glutamine-hydrolyzing] subunit A (EC 6.3.5.2) (Glutamine amidotransferase) [Gene: guaAA or Ta0944] - Thermoplasma acidophilum
1gpm	17.7	2.2	144	23	6.3.5.2	GMP synthase [glutamine-hydrolyzing] (EC 6.3.5.2) (Glutamine amidotransferase) (GMP synthetase) (GMPS) [Gene: guaA or b2507 or JW2491] - Escherichia coli (strain K12)
1wl8	16.8	2	130	26	6.3.5.2	GMP synthase [glutamine-hydrolyzing] subunit A (EC 6.3.5.2) (Glutamine amidotransferase) [Gene: guaAA or PH1346] - Pyrococcus horikoshii
1vcn	16.2	2.1	138	25	not found	CTP synthase (EC 6.3.4.2) (UTP--ammonia ligase) (CTP synthetase) [Gene: pyrG or TTHA1466] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1vco	16.1	2.3	143	24	6.3.4.2	CTP synthase (EC 6.3.4.2) (UTP--ammonia ligase) (CTP synthetase) [Gene: pyrG or TTHA1466] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1vcm	16	2.2	139	24	6.3.4.2	CTP synthase (EC 6.3.4.2) (UTP--ammonia ligase) (CTP synthetase) [Gene: pyrG or TTHA1466] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
2ad5	14.7	2.5	135	24	6.3.4.2	CTP synthase (EC 6.3.4.2) (UTP--ammonia ligase) (CTP synthetase) [Gene: pyrG or b2780 or JW2751] - Escherichia coli (strain K12)
1s1m	13.2	2.6	127	26	6.3.4.2	CTP synthase (EC 6.3.4.2) (UTP--ammonia ligase) (CTP synthetase) [Gene: pyrG or b2780 or JW2751] - Escherichia coli (strain K12)
1k9v	12.3	2.2	111	21	2.4.2.-	Imidazole glycerol phosphate synthase subunit hisH (EC 2.4.2.-) (IGP synthase glutamine amidotransferase subunit) (IGP synthase subunit hisH) (ImGP synthase subunit hisH) (IGPS subunit hisH) (TmHisH) [Gene: hisH or TM_1038] - Thermotoga maritima
1kxj	12	2.2	108	20	2.4.2.-	Imidazole glycerol phosphate synthase subunit hisH (EC 2.4.2.-) (IGP synthase glutamine amidotransferase subunit) (IGP synthase subunit hisH) (ImGP synthase subunit hisH) (IGPS subunit hisH) (TmHisH) [Gene: hisH or TM_1038] - Thermotoga maritima
1gpw	11.9	2.2	107	21	2.4.2.-	Imidazole glycerol phosphate synthase subunit hisH (EC 2.4.2.-) (IGP synthase glutamine amidotransferase subunit) (IGP synthase subunit hisH) (ImGP synthase subunit hisH) (IGPS subunit hisH) (TmHisH) [Gene: hisH or TM_1038] - Thermotoga maritima
1q7r	11.7	2.9	133	17	2.6.-.-	Glutamine amidotransferase subunit pdxT (EC 2.6.-.-) (Glutamine amidotransferase glutaminase subunit pdxT) [Gene: pdxT] - Bacillus stearothermophilus (Geobacillus stearothermophilus)
1l9x	11.5	2.3	119	21	3.4.19.9	Gamma-glutamyl hydrolase precursor (EC 3.4.19.9) (Gamma-Glu-X carboxypeptidase) (Conjugase) (GH) [Gene: GGH] - Homo sapiens (Human)
1r9g	11.5	2.5	119	16	2.6.-.-	Glutamine amidotransferase subunit pdxT (EC 2.6.-.-) (Glutamine amidotransferase glutaminase subunit pdxT) [Gene: pdxT or yaaE or BSU00120] - Bacillus subtilis
1o1y	10.8	2.2	112	25	not found	Crystal structure of putative glutamine amido transferase (TM1158) from Thermotoga maritima at 1.70 A resolution
1ox6	10.2	2.5	123	20	4.1.3.-	Imidazole glycerol phosphate synthase hisHF (IGP synthase) (ImGP synthase) (IGPS) [Includes: Glutamine amidotransferase (EC 2.4.2.-); Cyclase (EC 4.1.3.-)] [Gene: HIS7 or YBR248C or YBR1640] - Saccharomyces cerevisiae (Baker's yeast)
1ox4	10.2	2.6	123	20	4.1.3.-	Imidazole glycerol phosphate synthase hisHF (IGP synthase) (ImGP synthase) (IGPS) [Includes: Glutamine amidotransferase (EC 2.4.2.-); Cyclase (EC 4.1.3.-)] [Gene: HIS7 or YBR248C or YBR1640] - Saccharomyces cerevisiae (Baker's yeast)
2abw	10	2.8	118	13	not found	Glutaminase subunit of the plasmodial PLP synthase (Vitamin B6 biosynthesis)
1jvn	9.9	2.5	119	19	2.4.2.-	Imidazole glycerol phosphate synthase hisHF (IGP synthase) (ImGP synthase) (IGPS) [Includes: Glutamine amidotransferase (EC 2.4.2.-); Cyclase (EC 4.1.3.-)] [Gene: HIS7 or YBR248C or YBR1640] - Saccharomyces cerevisiae (Baker's yeast)
1ox5	9.9	2.6	122	19	4.1.3.-	Imidazole glycerol phosphate synthase hisHF (IGP synthase) (ImGP synthase) (IGPS) [Includes: Glutamine amidotransferase (EC 2.4.2.-); Cyclase (EC 4.1.3.-)] [Gene: HIS7 or YBR248C or YBR1640] - Saccharomyces cerevisiae (Baker's yeast)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1qdl	8.8	1.8	81	28	4.1.3.27	Anthranilate synthase component 1 (EC 4.1.3.27) (Anthranilate synthase component I) [Gene: trpE or SSO0893] - Sulfolobus solfataricus
1i7q	8.7	1.8	85	28	4.1.3.27	Anthranilate synthase component 1 (EC 4.1.3.27) (Anthranilate synthase component I) [Gene: trpE] - Serratia marcescens
1ox5	8	2.3	99	20	4.1.3.-	Imidazole glycerol phosphate synthase hisHF (IGP synthase) (ImGP synthase) (IGPS) [Includes: Glutamine amidotransferase (EC 2.4.2.-); Cyclase (EC 4.1.3.-)] [Gene: HIS7 or YBR248C or YBR1640] - Saccharomyces cerevisiae (Baker's yeast)
1chu	23.7	0			1.4.3.16	L-aspartate oxidase (EC 1.4.3.16) (LASPO) (Quinolinate synthetase B) [Gene: nadB or nicB or b2574 or JW2558] - Escherichia coli (strain K12)
1qlb	13.1	1.9	92	43	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or WS0831] - Wolinella succinogenes
2bs3	11.5	1.6	90	46	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or WS0831] - Wolinella succinogenes
1fum	11.5	1.9	89	58	not found	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or b4154 or JW4115] - Escherichia coli (strain K12)
1kfy	11.5	2	90	59	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or b4154 or JW4115] - Escherichia coli (strain K12)
1kf6	11.4	2	90	59	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or b4154 or JW4115] - Escherichia coli (strain K12)
2bs4	11.1	1.8	90	46	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or WS0831] - Wolinella succinogenes
1qla	11.1	2	91	44	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or WS0831] - Wolinella succinogenes
2bs4	10.9	2.2	90	46	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or WS0831] - Wolinella succinogenes
1l0v	10.9	3	92	57	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or b4154 or JW4115] - Escherichia coli (strain K12)
2b76	10.1	2.4	89	58	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or b4154 or JW4115] - Escherichia coli (strain K12)
1nen	9.5	2.4	91	47	1.3.99.1	Succinate dehydrogenase flavoprotein subunit (EC 1.3.99.1) [Gene: sdhA or b0723 or JW0713] - Escherichia coli (strain K12)
1nek	9.5	2.4	91	47	1.3.99.1	Succinate dehydrogenase flavoprotein subunit (EC 1.3.99.1) [Gene: sdhA or b0723 or JW0713] - Escherichia coli (strain K12)
1d4c	9.2	1.5	74	46	1.3.99.1	Fumarate reductase flavoprotein subunit precursor (EC 1.3.99.1) (Flavocytochrome c) (FL cyt) [Gene: SO_0970] - Shewanella oneidensis
1e7p	9.2	2.7	87	46	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or WS0831] - Wolinella succinogenes
2b76	8.8	2.2	84	60	not found	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) [Gene: frdA or b4154 or JW4115] - Escherichia coli (strain K12)
1qo8	8.6	1.6	78	36	1.3.99.1	Fumarate reductase flavoprotein subunit precursor (EC 1.3.99.1) (Iron(III)-induced flavocytochrome C3) (Ifc3) [Gene: ifcA or Sfri_2586] - Shewanella frigidimarina (strain NCIMB 400)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1qo8	8	1.6	70	37	1.3.99.1	Fumarate reductase flavoprotein subunit precursor (EC 1.3.99.1) (Iron(III)-induced flavocytochrome C3) (Ifc3) [Gene: ifcA or Sfri_2586] - Shewanella frigidimarina (strain NCIMB 400)
1d4c	7.9	1.7	74	46	1.3.99.1	Fumarate reductase flavoprotein subunit precursor (EC 1.3.99.1) (Flavocytochrome c) (FL cyt) [Gene: SO_0970] - Shewanella oneidensis
1d4d	7.2	2	74	45	1.3.99.1	Fumarate reductase flavoprotein subunit precursor (EC 1.3.99.1) (Flavocytochrome c) (FL cyt) [Gene: SO_0970] - Shewanella oneidensis
1d4c	6.8	2.3	74	46	1.3.99.1	Fumarate reductase flavoprotein subunit precursor (EC 1.3.99.1) (Flavocytochrome c) (FL cyt) [Gene: SO_0970] - Shewanella oneidensis
1ksu	6.4	2.4	74	42	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) (Flavocytochrome c) (Flavocytochrome c3) (Fcc3) [Gene: fccA or fcc3] - Shewanella frigidimarina
1kss	6.3	2.2	73	40	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) (Flavocytochrome c) (Flavocytochrome c3) (Fcc3) [Gene: fccA or fcc3] - Shewanella frigidimarina
1m64	6.1	2.2	72	38	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) (Flavocytochrome c) (Flavocytochrome c3) (Fcc3) [Gene: fccA or fcc3] - Shewanella frigidimarina
1y0p	6	2.2	74	41	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) (Flavocytochrome c) (Flavocytochrome c3) (Fcc3) [Gene: fccA or fcc3] - Shewanella frigidimarina
1qjd	5.8	2.3	75	40	1.3.99.1	Fumarate reductase flavoprotein subunit (EC 1.3.99.1) (Flavocytochrome c) (Flavocytochrome c3) (Fcc3) [Gene: fccA or fcc3] - Shewanella frigidimarina
1ctn	31.1	0			3.2.1.14	Chitinase A precursor (EC 3.2.1.14) [Gene: chiA] - Serratia marcescens
1eib	29.6	0.2	146	99	not found	CRYSTAL STRUCTURE OF CHITINASE A MUTANT D313A COMPLEXED WITH OCTA-N-ACETYLCHITOOCTAOSE (NAG)8.
1ffq	29.4	0.2	146	100	3.2.1.14	Chitinase A precursor (EC 3.2.1.14) [Gene: chiA] - Serratia marcescens
1ffr	29.3	0.2	146	99	3.2.1.14	Chitinase A precursor (EC 3.2.1.14) [Gene: chiA] - Serratia marcescens
1edq	29.3	0.3	146	100	not found	CRYSTAL STRUCTURE OF CHITINASE A FROM S. MARCESCENS AT 1.55 ANGSTROMS
1k9t	29.2	0.2	146	99	3.2.1.14	Chitinase A precursor (EC 3.2.1.14) [Gene: chiA] - Serratia marcescens
1x6n	27.9	0.3	146	100	3.2.1.14	Chitinase A precursor (EC 3.2.1.14) [Gene: chiA] - Serratia marcescens
1rd6	27.6	0.4	146	100	3.2.1.14	Chitinase A precursor (EC 3.2.1.14) [Gene: chiA] - Serratia marcescens
1x6l	27.1	0.3	143	100	3.2.1.14	Chitinase A precursor (EC 3.2.1.14) [Gene: chiA] - Serratia marcescens
1itx	23.1	1.2	137	35	3.2.1.14	Chitinase A1 precursor (EC 3.2.1.14) [Gene: chiA1] - Bacillus circulans
1ll4	21.7	0.9	126	40	3.2.1.14	Endochitinase 1 precursor (EC 3.2.1.14) (Complement-fixation antigen) (CF-antigen) (CF-AG) (CiX1) [Gene: CTS1] - Coccidioides posadasii
1d2k	21.7	0.9	126	41	3.2.1.14	Endochitinase 1 precursor (EC 3.2.1.14) (Complement-fixation antigen) (CF-antigen) (CF-AG) (CiX1) [Gene: CTS1] - Coccidioides posadasii
1hkk	21.4	1.1	132	33	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1kfw	21.1	1.3	135	34	not found	Structure of catalytic domain of psychrophilic chitinase B from Arthrobacter TAD20
1hkj	20.9	1.2	131	33	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
1wno	20.6	1	127	42	not found	Crystal structure of a native chitinase from Aspergillus fumigatus YJ-407
2a3b	20.6	1	127	42	not found	Crystal structure of Aspergillus fumigatus chitinase B1 in complex with caffeine
2a3a	20.6	1	127	42	not found	Crystal structure of Aspergillus fumigatus chitinase B1 in complex with theophylline
1hkm	20.6	1.1	130	34	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
2a3e	20.5	1	127	42	not found	Crystal structure of Aspergillus fumigatus chitinase B1 in complex with allosamidin
1wno	20.5	1	127	42	not found	Crystal structure of a native chitinase from Aspergillus fumigatus YJ-407
1lg2	20.3	1.1	132	33	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
1hki	20.3	1.2	129	34	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
1ll6	20.2	1	126	40	3.2.1.14	Endochitinase 1 precursor (EC 3.2.1.14) (Complement-fixation antigen) (CF-antigen) (CF-AG) (CiX1) [Gene: CTS1] - Coccidioides posadasii
1w9v	20.2	1	127	42	not found	SPECIFICITY AND AFFINITY OF NATURAL PRODUCT CYCLOPENTAPEPTIDE ARGIFIN AGAINST ASPERGILLUS FUMIGATUS
1ll7	20.1	0.9	126	40	3.2.1.14	Endochitinase 1 precursor (EC 3.2.1.14) (Complement-fixation antigen) (CF-antigen) (CF-AG) (CiX1) [Gene: CTS1] - Coccidioides posadasii
1guv	20.1	1.2	132	33	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
1e9l	20.1	1.2	129	30	0.0.0.0	Chitinase-3-like protein 3 precursor (Secretory protein Ym1) (Eosinophil chemotactic cytokine) (ECF-L) [Gene: Chi3l3 or Ym1] - Mus musculus (Mouse)
1w9u	19.8	1	127	42	not found	SPECIFICITY AND AFFNITY OF NATURAL PRODUCT CYCLOPENTAPEPTIDE INHIBITOR ARGADIN AGAINST ASPERGILLUS FUMIGATUS CHITINASE
1w9p	19.8	1	127	42	not found	SPECIFICITY AND AFFINITY OF NATURAL PRODUCT CYCLOPENTAPEPTIDE INHIBITORS AGAINST ASPERGILLUS FUMIGATUS, HUMAN AND BACTERIAL CHITINASEFRA
1ehn	19	0.2	112	99	not found	CRYSTAL STRUCTURE OF CHITINASE A MUTANT E315Q COMPLEXED WITH OCTA-N-ACETYLCHITOOCTAOSE (NAG)8.
1wb0	19	1.2	132	33	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
1waw	18.9	1.1	132	33	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
1e6z	17.8	1.3	125	34	not found	CHITINASE B FROM SERRATIA MARCESCENS WILDTYPE IN COMPLEX WITH CATALYTIC INTERMEDIATE
1w1p	17.7	1.6	126	34	not found	CRYSTAL STRUCTURE OF S. MARCESCENS CHITINASE B IN COMPLEX WITH THE CYCLIC DIPEPTIDE INHIBITOR CYCLO-(GLY-L-PRO) AT 2.1 A RESOLUTION
1nh6	17.6	0.2	111	99	3.2.1.14	Chitinase A precursor (EC 3.2.1.14) [Gene: chiA] - Serratia marcescens

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1w1y	17.6	1.6	126	34	not found	CRYSTAL STRUCTURE OF S. MARCESCENS CHITINASE B IN COMPLEX WITH THE CYCLIC DIPEPTIDE INHIBITOR CYCLO-(L-TYR-L-PRO) AT 1.85 A RESOLUTION
1.00E+15	17.5	1.3	118	36	not found	CHITINASE B FROM SERRATIA MARCESCENS
1e6r	17.5	1.3	125	34	not found	CHITINASE B FROM SERRATIA MARCESCENS WILDTYPE IN COMPLEX WITH INHIBITOR ALLOSAMIDIN
1goi	17.5	1.3	117	36	not found	CRYSTAL STRUCTURE OF THE D140N MUTANT OF CHITINASE B FROM SERRATIA MARCESCENS AT 1.45 A RESOLUTION
1ur8	17.5	1.6	126	34	not found	INTERACTIONS OF A FAMILY 18 CHITINASE WITH THE DESIGNED INHIBITOR HM508, AND ITS DEGRADATION PRODUCT, CHITOBIONO-DELTA-LACTONE
1ogg	17.5	1.6	126	33	3.2.1.14	Chitinase B precursor (EC 3.2.1.14) [Gene: chiB] - Serratia marcescens
1gpf	17.4	1.3	117	37	not found	CHITINASE B FROM SERRATIA MARCESCENS IN COMPLEX WITH INHIBITOR PSAMMAPLIN
1ogb	17.3	1.6	123	34	3.2.1.14	Chitinase B precursor (EC 3.2.1.14) [Gene: chiB] - Serratia marcescens
1o6i	17.3	1.6	123	35	3.2.1.14	Chitinase B precursor (EC 3.2.1.14) [Gene: chiB] - Serratia marcescens
1w1v	17.2	1.6	123	35	not found	CRYSTAL STRUCTURE OF S. MARCESCENS CHITINASE B IN COMPLEX WITH THE CYCLIC DIPEPTIDE INHIBITOR CYCLO-(L-ARG-L-PRO) AT 1.85 A RESOLUTION
1w1t	17.2	1.6	123	35	not found	CRYSTAL STRUCTURE OF S. MARCESCENS CHITINASE B IN COMPLEX WITH THE CYCLIC DIPEPTIDE INHIBITOR CYCLO-(HIS-L-PRO) AT 1.9 A RESOLUTION
1ur9	17	1.6	123	34	not found	INTERACTIONS OF A FAMILY 18 CHITINASE WITH THE DESIGNED INHIBITOR HM508, AND ITS DEGRADATION PRODUCT, CHITOBIONO-DELTA-LACTONE
1h0g	16.9	1.4	117	37	3.2.1.14	Chitinase B precursor (EC 3.2.1.14) [Gene: chiB] - Serratia marcescens
1h0i	16.9	1.7	120	36	3.2.1.14	Chitinase B precursor (EC 3.2.1.14) [Gene: chiB] - Serratia marcescens
1e6p	16.7	1.3	118	36	not found	CHITINASE B FROM SERRATIA MARCESCENS INACTIVE MUTANT E144Q
1e6n	16.4	1.4	118	36	not found	CHITINASE B FROM SERRATIA MARCESCENS INACTIVE MUTANT E144Q IN COMPLEX WITH N-ACETYLGLUCOSAMINE-PENTAMER
1vf8	16.3	1.2	119	30	0.0.0.0	Chitinase-3-like protein 3 precursor (Secretory protein Ym1) (Eosinophil chemotactic cytokine) (ECF-L) [Gene: Chi3l3 or Ym1] - Mus musculus (Mouse)
1lg2	15.4	1.5	113	35	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
1lg1	15.2	1.5	113	35	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
2a3c	14.9	0.9	95	47	not found	Crystal structure of Aspergillus fumigatus chitinase B1 in complex with pentoxifylline
1wb0	14.4	1.7	109	35	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
1waw	14.2	1.7	109	35	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1w9u	13.5	2.1	104	41	not found	SPECIFICITY AND AFFNITY OF NATURAL PRODUCT CYCLOPENTAPEPTIDE INHIBITOR ARGADIN AGAINST ASPERGILLUS FUMIGATUS CHITINASE
1w9p	13.4	2.1	104	41	not found	SPECIFICITY AND AFFINITY OF NATURAL PRODUCT CYCLOPENTAPEPTIDE INHIBITORS AGAINST ASPERGILLUS FUMIGATUS, HUMAN AND BACTERIAL CHITINASEFRA
1w9v	13.4	2.4	103	43	not found	SPECIFICITY AND AFFINITY OF NATURAL PRODUCT CYCLOPENTAPEPTIDE ARGIFIN AGAINST ASPERGILLUS FUMIGATUS
1wno	13.3	1.7	103	43	not found	Crystal structure of a native chitinase from Aspergillus fumigatus YJ-407
1h0g	13.3	1.8	105	35	3.2.1.14	Chitinase B precursor (EC 3.2.1.14) [Gene: chiB] - Serratia marcescens
1lq0	12.9	1.4	98	35	3.2.1.14	Chitotriosidase-1 precursor (EC 3.2.1.14) (Chitinase-1) [Gene: CHIT1] - Homo sapiens (Human)
1ur9	11.7	1.5	100	35	not found	INTERACTIONS OF A FAMILY 18 CHITINASE WITH THE DESIGNED INHIBITOR HM508, AND ITS DEGRADATION PRODUCT, CHITOBIONO-DELTA-LACTONE
1w1p	11.4	1.5	97	36	not found	CRYSTAL STRUCTURE OF S. MARCESCENS CHITINASE B IN COMPLEX WITH THE CYCLIC DIPEPTIDE INHIBITOR CYCLO-(GLY-L-PRO) AT 2.1 A RESOLUTION
1w1v	11.4	1.5	97	36	not found	CRYSTAL STRUCTURE OF S. MARCESCENS CHITINASE B IN COMPLEX WITH THE CYCLIC DIPEPTIDE INHIBITOR CYCLO-(L-ARG-L-PRO) AT 1.85 A RESOLUTION
1w1y	11.2	1.5	97	36	not found	CRYSTAL STRUCTURE OF S. MARCESCENS CHITINASE B IN COMPLEX WITH THE CYCLIC DIPEPTIDE INHIBITOR CYCLO-(L-TYR-L-PRO) AT 1.85 A RESOLUTION
1djq	7	3.5	117	9	1.5.8.2	Trimethylamine dehydrogenase (EC 1.5.8.2) (TMADh) [Gene: tmd] - Methylophilus methylotrophus (Bacterium W3A1)
1o94	6.6	3.3	113	9	1.5.8.2	Trimethylamine dehydrogenase (EC 1.5.8.2) (TMADh) [Gene: tmd] - Methylophilus methylotrophus (Bacterium W3A1)
1f76	5.6	3.8	105	10	1.3.3.1	Dihydroorotate dehydrogenase (EC 1.3.3.1) (Dihydroorotate oxidase) (DHOdehase) (DHODase) (DHOD) [Gene: pyrD or b0945 or JW0928] - Escherichia coli (strain K12)
1oif	5.5	3	93	11	3.2.1.21	Beta-glucosidase A (EC 3.2.1.21) (Gentiobiase) (Cellobiase) (Beta-D-glucoside glucohydrolase) [Gene: bglA] - Thermotoga maritima
1o7a	5.2	3.1	89	8	3.2.1.52	Beta-hexosaminidase beta chain precursor (EC 3.2.1.52) (N-acetyl-beta-glucosaminidase) (Beta-N-acetylhexosaminidase) (Hexosaminidase B) (Cervical cancer proto-oncogene 7 protein) (HCC-7) [Contains: Beta-hexosaminidase beta-B chain; Beta-hexosaminidase beta-A chain] [Gene: HEXB or HCC7] - Homo sapiens (Human)
1ktb	5.2	3.6	99	8	3.2.1.49	Alpha-N-acetylgalactosaminidase (EC 3.2.1.49) (Alpha-galactosidase B) [Gene: NAGA] - Gallus gallus (Chicken)
1r47	5.2	3.6	102	12	3.2.1.22	Alpha-galactosidase A precursor (EC 3.2.1.22) (Melibiase) (Alpha-D-galactoside galactohydrolase) (Alpha-D-galactosidase A) (Agalsidase) [Gene: GLA] - Homo sapiens (Human)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1r46	5.2	3.7	103	12	3.2.1.22	Alpha-galactosidase A precursor (EC 3.2.1.22) (Melibiase) (Alpha-D-galactoside galactohydrolase) (Alpha-D-galactosidase A) (Agalsidase) [Gene: GLA] - Homo sapiens (Human)
1np0	4.7	3.4	91	8	3.2.1.52	Beta-hexosaminidase beta chain precursor (EC 3.2.1.52) (N-acetyl-beta-glucosaminidase) (Beta-N-acetylhexosaminidase) (Hexosaminidase B) (Cervical cancer proto-oncogene 7 protein) (HCC-7) [Contains: Beta-hexosaminidase beta-B chain; Beta-hexosaminidase beta-A chain] [Gene: HEXB or HCC7] - Homo sapiens (Human)
1nou	4.7	3.5	91	8	3.2.1.52	Beta-hexosaminidase beta chain precursor (EC 3.2.1.52) (N-acetyl-beta-glucosaminidase) (Beta-N-acetylhexosaminidase) (Hexosaminidase B) (Cervical cancer proto-oncogene 7 protein) (HCC-7) [Contains: Beta-hexosaminidase beta-B chain; Beta-hexosaminidase beta-A chain] [Gene: HEXB or HCC7] - Homo sapiens (Human)
1now	4.5	3.5	91	8	3.2.1.52	Beta-hexosaminidase beta chain precursor (EC 3.2.1.52) (N-acetyl-beta-glucosaminidase) (Beta-N-acetylhexosaminidase) (Hexosaminidase B) (Cervical cancer proto-oncogene 7 protein) (HCC-7) [Contains: Beta-hexosaminidase beta-B chain; Beta-hexosaminidase beta-A chain] [Gene: HEXB or HCC7] - Homo sapiens (Human)
1e5n	4.1	3	89	10	3.2.1.8	Endo-1,4-beta-xylanase A precursor (EC 3.2.1.8) (Xylanase A) (1,4-beta-D-xylan xylanohydrolase A) (XYLA) [Gene: xynA] - Pseudomonas fluorescens
1dgj	27.8	0			not found	CRYSTAL STRUCTURE OF THE ALDEHYDE OXIDOREDUCTASE FROM DESULFOVIBRIO DESULFURICANS ATCC 27774
1hlr	22.5	0.4	114	80	not found	Aldehyde oxidoreductase (EC 1.2.99.7) (Molybdenum iron sulfur protein) [Gene: mop] - Desulfovibrio gigas
1fo4	18.2	1.3	110	38	1.17.3.2	Xanthine dehydrogenase/oxidase [Includes: Xanthine dehydrogenase (EC 1.17.1.4) (XD); Xanthine oxidase (EC 1.17.3.2) (XO) (Xanthine oxidoreductase)] [Gene: XDH] - Bos taurus (Bovine)
1wyg	17.8	1.3	109	38	1.17.3.2	Xanthine dehydrogenase/oxidase [Includes: Xanthine dehydrogenase (EC 1.17.1.4) (XD); Xanthine oxidase (EC 1.17.3.2) (XO) (Xanthine oxidoreductase)] [Gene: Xdh] - Rattus norvegicus (Rat)
1fiq	17.1	1.1	105	37	1.17.1.4	Xanthine dehydrogenase/oxidase [Includes: Xanthine dehydrogenase (EC 1.17.1.4) (XD); Xanthine oxidase (EC 1.17.3.2) (XO) (Xanthine oxidoreductase)] [Gene: XDH] - Bos taurus (Bovine)
1jro	17.1	1.2	106	33	not found	Crystal Structure of Xanthine Dehydrogenase from Rhodobacter capsulatus
1jrp	17	1.2	106	33	not found	Crystal Structure of Xanthine Dehydrogenase inhibited by alloxanthine from Rhodobacter capsulatus
1n5x	17	2.2	109	39	1.17.1.4	Xanthine dehydrogenase/oxidase [Includes: Xanthine dehydrogenase (EC 1.17.1.4) (XD); Xanthine oxidase (EC 1.17.3.2) (XO) (Xanthine oxidoreductase)] [Gene: XDH] - Bos taurus (Bovine)
1n62	13.9	1.5	102	31	1.2.99.2	Carbon monoxide dehydrogenase large chain (EC 1.2.99.2) (CO dehydrogenase subunit L) (CO-DH L) [Gene: coxL] - Oligotropha carboxidovorans (Pseudomonas carboxydovorans)
1ffu	13.4	1.6	97	32	1.2.99.2	Carbon monoxide dehydrogenase large chain (EC 1.2.99.2) (CO dehydrogenase subunit L) (CO-DH L) [Gene: cutL] - Hydrogenophaga pseudoflava (Pseudomonas carboxydoflava)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1n61	13.3	1.5	98	32	1.2.99.2	Carbon monoxide dehydrogenase large chain (EC 1.2.99.2) (CO dehydrogenase subunit L) (CO-DH L) [Gene: coxL] - Oligotropha carboxidovorans (Pseudomonas carboxydovorans)
1n63	13.3	1.5	98	32	1.2.99.2	Carbon monoxide dehydrogenase large chain (EC 1.2.99.2) (CO dehydrogenase subunit L) (CO-DH L) [Gene: coxL] - Oligotropha carboxidovorans (Pseudomonas carboxydovorans)
1n60	13.2	1.5	98	32	1.2.99.2	Carbon monoxide dehydrogenase large chain (EC 1.2.99.2) (CO dehydrogenase subunit L) (CO-DH L) [Gene: coxL] - Oligotropha carboxidovorans (Pseudomonas carboxydovorans)
1n5w	13	1.9	99	32	1.2.99.2	Carbon monoxide dehydrogenase large chain (EC 1.2.99.2) (CO dehydrogenase subunit L) (CO-DH L) [Gene: coxL] - Oligotropha carboxidovorans (Pseudomonas carboxydovorans)
1qj2	12.7	1.4	96	32	not found	Carbon monoxide dehydrogenase large chain (EC 1.2.99.2) (CO dehydrogenase subunit L) (CO-DH L) [Gene: coxL] - Oligotropha carboxidovorans (Pseudomonas carboxydovorans)
1ffv	12.3	2	97	34	1.2.99.2	Carbon monoxide dehydrogenase large chain (EC 1.2.99.2) (CO dehydrogenase subunit L) (CO-DH L) [Gene: cutL] - Hydrogenophaga pseudoflava (Pseudomonas carboxydoflava)
1jrp	7	1.3	72	36	not found	Crystal Structure of Xanthine Dehydrogenase inhibited by alloxanthine from Rhodobacter capsulatus
1sb3	5.8	1.8	69	33	1.3.99.20	4-hydroxybenzoyl-CoA reductase subunit alpha (EC 1.3.99.20) [Gene: hcrA] - Thauera aromatica
1rm6	5.8	3.3	70	36	1.3.99.20	4-hydroxybenzoyl-CoA reductase subunit alpha (EC 1.3.99.20) [Gene: hcrA] - Thauera aromatica
1dgp	33.3	0			4.2.3.9	Aristolochene synthase (EC 4.2.3.9) (Sesquiterpene cyclase) (AS) [Gene: ARI1] - Penicillium roqueforti
1di1	31.6	0.1	165	100	4.2.3.9	Aristolochene synthase (EC 4.2.3.9) (Sesquiterpene cyclase) (AS) [Gene: ARI1] - Penicillium roqueforti
1f1n	28.9	0.3	161	100	not found	Aristolochene synthase (EC 4.2.3.9) (Sesquiterpene cyclase) (AS) [Gene: ARI1] - Penicillium roqueforti
1f1k	28.8	0.3	161	100	not found	Aristolochene synthase (EC 4.2.3.9) (Sesquiterpene cyclase) (AS) [Gene: ARI1] - Penicillium roqueforti
1hm7	16.4	2.7	154	16	4.2.3.7	Pentalenene synthase (EC 4.2.3.7) (PS) (Sesquiterpene synthase) (Sesquiterpene cyclase) - Streptomyces sp. (strain UC5319)
1ps1	13.3	2.8	139	17	4.2.3.7	Pentalenene synthase (EC 4.2.3.7) (PS) (Sesquiterpene synthase) (Sesquiterpene cyclase) - Streptomyces sp. (strain UC5319)
1hm4	13	2.5	127	16	4.2.3.7	Pentalenene synthase (EC 4.2.3.7) (PS) (Sesquiterpene synthase) (Sesquiterpene cyclase) - Streptomyces sp. (strain UC5319)
1hm4	10.6	2.3	104	17	4.2.3.7	Pentalenene synthase (EC 4.2.3.7) (PS) (Sesquiterpene synthase) (Sesquiterpene cyclase) - Streptomyces sp. (strain UC5319)
5eau	7.9	2.6	104	10	4.2.3.9	Aristolochene synthase (EC 4.2.3.9) (5-epi-aristolochene synthase) (EAS) - Nicotiana tabacum (Common tobacco)
1hxa	7.4	2.6	107	9	4.2.3.9	Aristolochene synthase (EC 4.2.3.9) (5-epi-aristolochene synthase) (EAS) - Nicotiana tabacum (Common tobacco)
1hxc	7.4	2.7	105	10	4.2.3.9	Aristolochene synthase (EC 4.2.3.9) (5-epi-aristolochene synthase) (EAS) - Nicotiana tabacum (Common tobacco)
5eat	7.2	2.8	104	11	4.2.3.9	Aristolochene synthase (EC 4.2.3.9) (5-epi-aristolochene synthase) (EAS) - Nicotiana tabacum (Common tobacco)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1n24	5.8	3.2	106	8	5.5.1.8	(+)-bornyl diphosphate synthase, chloroplast precursor (EC 5.5.1.8) (SBS) (BPPS) - Salvia officinalis (Sage)
1n20	5.6	3.2	106	8	5.5.1.8	(+)-bornyl diphosphate synthase, chloroplast precursor (EC 5.5.1.8) (SBS) (BPPS) - Salvia officinalis (Sage)
1n23	5.6	3.2	106	8	5.5.1.8	(+)-bornyl diphosphate synthase, chloroplast precursor (EC 5.5.1.8) (SBS) (BPPS) - Salvia officinalis (Sage)
1n1b	5.4	3.2	102	8	5.5.1.8	(+)-bornyl diphosphate synthase, chloroplast precursor (EC 5.5.1.8) (SBS) (BPPS) - Salvia officinalis (Sage)
1n1z	4.9	3.2	99	8	5.5.1.8	(+)-bornyl diphosphate synthase, chloroplast precursor (EC 5.5.1.8) (SBS) (BPPS) - Salvia officinalis (Sage)
1n22	4.8	3.2	99	8	5.5.1.8	(+)-bornyl diphosphate synthase, chloroplast precursor (EC 5.5.1.8) (SBS) (BPPS) - Salvia officinalis (Sage)
1dia	21.5	0			1.5.1.5	C-1-tetrahydrofolate synthase, cytoplasmic (C1-THF synthase) [Includes: Methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5); Methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9); Formyltetrahydrofolate synthetase (EC 6.3.4.3)] [Gene: MTHFD1 or MTHFC or MTHFD] - Homo sapiens (Human)
1dig	20.8	0.1	90	100	6.3.4.3	C-1-tetrahydrofolate synthase, cytoplasmic (C1-THF synthase) [Includes: Methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5); Methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9); Formyltetrahydrofolate synthetase (EC 6.3.4.3)] [Gene: MTHFD1 or MTHFC or MTHFD] - Homo sapiens (Human)
2c2x	13.7	1.3	87	36	1.5.1.5	Bifunctional protein folD [Includes: Methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5); Methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9)] [Gene: folD or Rv3356c or MT3464] - Mycobacterium tuberculosis
1ee9	11.4	1.3	82	22	1.5.1.15	Methylenetetrahydrofolate dehydrogenase [NAD+] (EC 1.5.1.15) [Gene: MTD1 or YKR080W or YKR400] - Saccharomyces cerevisiae (Baker's yeast)
1ekx	4.7	2.6	68	9	2.1.3.2	Aspartate carbamoyltransferase catalytic chain (EC 2.1.3.2) (Aspartate transcarbamylase) (ATCase) [Gene: pyrB or b4245 or JW4204] - Escherichia coli (strain K12)
1ekx	4.6	2.6	68	9	2.1.3.2	Aspartate carbamoyltransferase catalytic chain (EC 2.1.3.2) (Aspartate transcarbamylase) (ATCase) [Gene: pyrB or b4245 or JW4204] - Escherichia coli (strain K12)
8atc	4.5	2.6	67	9	2.1.3.2	Aspartate carbamoyltransferase catalytic chain (EC 2.1.3.2) (Aspartate transcarbamylase) (ATCase) [Gene: pyrB or b4245 or JW4204] - Escherichia coli (strain K12)
1dtw	17.6	0			1.2.4.4	2-oxoisovalerate dehydrogenase subunit alpha, mitochondrial precursor (EC 1.2.4.4) (Branched-chain alpha-keto acid dehydrogenase E1 component alpha chain) (BCKDH E1-alpha) (BCKDE1A) [Gene: BCKDHA] - Homo sapiens (Human)
1pvd	4.8	2.4	58	10	4.1.1.1	Pyruvate decarboxylase isozyme 1 (EC 4.1.1.1) (EC 4.1.1.-) [Gene: PDC1 or YLR044C or L2104] - Saccharomyces cerevisiae (Baker's yeast)
1pyd	4.7	2.3	57	9	4.1.1.1	Pyruvate decarboxylase isozyme 1 (EC 4.1.1.1) (EC 4.1.1.-) [Gene: PDC1 or YLR044C or L2104] - Saccharomyces cerevisiae (Baker's yeast)
1qpb	4.7	2.3	56	9	4.1.1.1	Pyruvate decarboxylase isozyme 1 (EC 4.1.1.1) (EC 4.1.1.-) [Gene: PDC1 or YLR044C or L2104] - Saccharomyces cerevisiae (Baker's yeast)
1ovm	4.5	2.7	59	8	4.1.1.74	Indole-3-pyruvate decarboxylase (EC 4.1.1.74) (Indolepyruvate decarboxylase) [Gene: ipdC] - Enterobacter cloacae

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1e2t	38	0			2.3.1.118	N-hydroxyarylamine O-acetyltransferase (EC 2.3.1.118) (Arylhydroxamate N,O-acetyltransferase) (Arylamine N-acetyltransferase) (NAT101) [Gene: nhoA or STM1582] - Salmonella typhimurium
1w6f	29.7	1.1	178	43	2.3.1.5	Arylamine N-acetyltransferase (EC 2.3.1.5) [Gene: nat] - Mycobacterium smegmatis
2bsz	29.7	1.2	182	43	not found	STRUCTURE OF MESORHIZOBIUM LOTI ARYLAMINE N-ACETYLTRANSFERASE 1
1gx3	29.1	1.1	177	44	2.3.1.5	Arylamine N-acetyltransferase (EC 2.3.1.5) [Gene: nat] - Mycobacterium smegmatis
1w4t	26.3	1.2	168	35	not found	X-RAY CRYSTALLOGRAPHIC STRUCTURE OF PSEUDOMONAS AERUGINOSA ARYLAMINE N-ACETYLTRANSFERASE
1ex0	8.1	3.1	115	13	2.3.2.13	Coagulation factor XIII A chain precursor (EC 2.3.2.13) (Coagulation factor XIIIa) (Protein-glutamine gamma-glutamyltransferase A chain) (Transglutaminase A chain) [Gene: F13A1 or F13A] - Homo sapiens (Human)
1sgx	7.9	3.4	121	14	2.3.2.13	Protein-glutamine gamma-glutamyltransferase E precursor (EC 2.3.2.13) (TGase E) (TGE) (TG(E)) (Transglutaminase-3) [Contains: Protein-glutamine gamma-glutamyltransferase E 50 kDa non-catalytic chain; Protein-glutamine gamma-glutamyltransferase E 27 kDa catalytic chain] [Gene: TGM3] - Homo sapiens (Human)
1f13	7.8	3.2	117	14	2.3.2.13	Coagulation factor XIII A chain precursor (EC 2.3.2.13) (Coagulation factor XIIIa) (Protein-glutamine gamma-glutamyltransferase A chain) (Transglutaminase A chain) [Gene: F13A1 or F13A] - Homo sapiens (Human)
1fie	7.8	3.5	118	13	2.3.2.13	Coagulation factor XIII A chain precursor (EC 2.3.2.13) (Coagulation factor XIIIa) (Protein-glutamine gamma-glutamyltransferase A chain) (Transglutaminase A chain) [Gene: F13A1 or F13A] - Homo sapiens (Human)
1vjj	7.7	3.3	116	13	2.3.2.13	Protein-glutamine gamma-glutamyltransferase E precursor (EC 2.3.2.13) (TGase E) (TGE) (TG(E)) (Transglutaminase-3) [Contains: Protein-glutamine gamma-glutamyltransferase E 50 kDa non-catalytic chain; Protein-glutamine gamma-glutamyltransferase E 27 kDa catalytic chain] [Gene: TGM3] - Homo sapiens (Human)
1ggu	7.6	3.5	121	12	2.3.2.13	Coagulation factor XIII A chain precursor (EC 2.3.2.13) (Coagulation factor XIIIa) (Protein-glutamine gamma-glutamyltransferase A chain) (Transglutaminase A chain) [Gene: F13A1 or F13A] - Homo sapiens (Human)
1qrk	7.6	3.5	115	13	2.3.2.13	Coagulation factor XIII A chain precursor (EC 2.3.2.13) (Coagulation factor XIIIa) (Protein-glutamine gamma-glutamyltransferase A chain) (Transglutaminase A chain) [Gene: F13A1 or F13A] - Homo sapiens (Human)
1nug	7.4	3.2	115	13	2.3.2.13	Protein-glutamine gamma-glutamyltransferase E precursor (EC 2.3.2.13) (TGase E) (TGE) (TG(E)) (Transglutaminase-3) [Contains: Protein-glutamine gamma-glutamyltransferase E 50 kDa non-catalytic chain; Protein-glutamine gamma-glutamyltransferase E 27 kDa catalytic chain] [Gene: TGM3] - Homo sapiens (Human)
1l9m	7.4	3.3	112	13	2.3.2.13	Protein-glutamine gamma-glutamyltransferase E precursor (EC 2.3.2.13) (TGase E) (TGE) (TG(E)) (Transglutaminase-3) [Contains: Protein-glutamine gamma-glutamyltransferase E 50 kDa non-catalytic chain; Protein-glutamine gamma-glutamyltransferase E 27 kDa catalytic chain] [Gene: TGM3] - Homo sapiens (Human)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1ggt	7.4	3.4	114	15	2.3.2.13	Coagulation factor XIII A chain precursor (EC 2.3.2.13) (Coagulation factor XIIIa) (Protein-glutamine gamma-glutamyltransferase A chain) (Transglutaminase A chain) [Gene: F13A1 or F13A] - Homo sapiens (Human)
1kv3	7.4	3.8	122	9	2.3.2.13	Protein-glutamine gamma-glutamyltransferase 2 (EC 2.3.2.13) (Tissue transglutaminase) (TGase C) (TGC) (TG(C)) (Transglutaminase-2) (TGase-H) [Gene: TGM2] - Homo sapiens (Human)
1l9n	7.3	3.2	110	14	2.3.2.13	Protein-glutamine gamma-glutamyltransferase E precursor (EC 2.3.2.13) (TGase E) (TGE) (TG(E)) (Transglutaminase-3) [Contains: Protein-glutamine gamma-glutamyltransferase E 50 kDa non-catalytic chain; Protein-glutamine gamma-glutamyltransferase E 27 kDa catalytic chain] [Gene: TGM3] - Homo sapiens (Human)
1ggy	7.3	3.3	116	14	2.3.2.13	Coagulation factor XIII A chain precursor (EC 2.3.2.13) (Coagulation factor XIIIa) (Protein-glutamine gamma-glutamyltransferase A chain) (Transglutaminase A chain) [Gene: F13A1 or F13A] - Homo sapiens (Human)
1rle	7.3	3.4	116	13	2.3.2.13	Protein-glutamine gamma-glutamyltransferase E precursor (EC 2.3.2.13) (TGase E) (TGE) (TG(E)) (Transglutaminase-3) [Contains: Protein-glutamine gamma-glutamyltransferase E 50 kDa non-catalytic chain; Protein-glutamine gamma-glutamyltransferase E 27 kDa catalytic chain] [Gene: TGM3] - Homo sapiens (Human)
1nud	7.1	3.3	120	14	2.3.2.13	Protein-glutamine gamma-glutamyltransferase E precursor (EC 2.3.2.13) (TGase E) (TGE) (TG(E)) (Transglutaminase-3) [Contains: Protein-glutamine gamma-glutamyltransferase E 50 kDa non-catalytic chain; Protein-glutamine gamma-glutamyltransferase E 27 kDa catalytic chain] [Gene: TGM3] - Homo sapiens (Human)
1evu	7	3.4	115	14	2.3.2.13	Coagulation factor XIII A chain precursor (EC 2.3.2.13) (Coagulation factor XIIIa) (Protein-glutamine gamma-glutamyltransferase A chain) (Transglutaminase A chain) [Gene: F13A1 or F13A] - Homo sapiens (Human)
1rll	7	3.7	127	13	not found	Protein-glutamine gamma-glutamyltransferase E precursor (EC 2.3.2.13) (TGase E) (TGE) (TG(E)) (Transglutaminase-3) [Contains: Protein-glutamine gamma-glutamyltransferase E 50 kDa non-catalytic chain; Protein-glutamine gamma-glutamyltransferase E 27 kDa catalytic chain] [Gene: TGM3] - Homo sapiens (Human)
1nuf	6.9	3.4	116	15	2.3.2.13	Protein-glutamine gamma-glutamyltransferase E precursor (EC 2.3.2.13) (TGase E) (TGE) (TG(E)) (Transglutaminase-3) [Contains: Protein-glutamine gamma-glutamyltransferase E 50 kDa non-catalytic chain; Protein-glutamine gamma-glutamyltransferase E 27 kDa catalytic chain] [Gene: TGM3] - Homo sapiens (Human)
1e8c	34.4	0			6.3.2.13	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2,6-diaminopimelate ligase (EC 6.3.2.13) (UDP-MurNAc-L-Ala-D-Glu:meso-diaminopimelate ligase) (Meso-diaminopimelate-adding enzyme) (Meso-A2pm-adding enzyme) (UDP-N-acetylmuramyl-tripeptide synthetase) (UDP-MurNAc-tripeptide synthetase) [Gene: murE or b0085 or JW0083] - Escherichia coli (strain K12)
2uag	15.6	2	135	32	6.3.2.9	UDP-N-acetylmuramoylalanine--D-glutamate ligase (EC 6.3.2.9) (UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase) (D-glutamic acid-adding enzyme) [Gene: murD or b0088 or JW0086] - Escherichia coli (strain K12)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1w78	12.9	2.4	132	27	6.3.2.17	Bifunctional protein folC [Includes: Folylpolyglutamate synthase (EC 6.3.2.17) (Folylpoly-gamma-glutamate synthetase) (FPGS) (Tetrahydrofolate synthase) (Tetrahydrofolylpolyglutamate synthase); Dihydrofolate synthase (EC 6.3.2.12)] [Gene: folC or dedC or b2315 or JW2312] - Escherichia coli (strain K12)
1p31	12.7	2.4	137	28	6.3.2.8	UDP-N-acetylmuramate--L-alanine ligase (EC 6.3.2.8) (UDP-N-acetylmuramoyl-L-alanine synthetase) [Gene: murC or HI1139] - Haemophilus influenzae
1j6u	11.9	2.7	137	28	6.3.2.8	UDP-N-acetylmuramate--L-alanine ligase (EC 6.3.2.8) (UDP-N-acetylmuramoyl-L-alanine synthetase) [Gene: murC or TM_0231] - Thermotoga maritima
1p3d	11.4	2.2	125	29	6.3.2.8	UDP-N-acetylmuramate--L-alanine ligase (EC 6.3.2.8) (UDP-N-acetylmuramoyl-L-alanine synthetase) [Gene: murC or HI1139] - Haemophilus influenzae
1eeh	11.3	1.9	112	33	6.3.2.9	UDP-N-acetylmuramoylalanine--D-glutamate ligase (EC 6.3.2.9) (UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase) (D-glutamic acid-adding enzyme) [Gene: murD or b0088 or JW0086] - Escherichia coli (strain K12)
1jbv	11.2	2.5	126	17	6.3.2.17	Folylpolyglutamate synthase (EC 6.3.2.17) (Folylpoly-gamma-glutamate synthetase) (FPGS) (Tetrahydrofolate synthase) (Tetrahydrofolylpolyglutamate synthase) [Gene: fgs] - Lactobacillus casei
1fgs	10.4	2.4	121	16	6.3.2.17	Folylpolyglutamate synthase (EC 6.3.2.17) (Folylpoly-gamma-glutamate synthetase) (FPGS) (Tetrahydrofolate synthase) (Tetrahydrofolylpolyglutamate synthase) [Gene: fgs] - Lactobacillus casei
1gqq	9.1	2.1	94	27	6.3.2.8	UDP-N-acetylmuramate--L-alanine ligase (EC 6.3.2.8) (UDP-N-acetylmuramoyl-L-alanine synthetase) [Gene: murC or HI1139] - Haemophilus influenzae
1uag	6.7	2	83	34	6.3.2.9	UDP-N-acetylmuramoylalanine--D-glutamate ligase (EC 6.3.2.9) (UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase) (D-glutamic acid-adding enzyme) [Gene: murD or b0088 or JW0086] - Escherichia coli (strain K12)
1gg4	6.6	2.5	93	19	6.3.2.10	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase (EC 6.3.2.10) (UDP-MurNAc-pentapeptide synthetase) (D-alanyl-D-alanine-adding enzyme) [Gene: murF or mra or b0086 or JW0084] - Escherichia coli (strain K12)
1ea0	32	0			1.4.1.13	Glutamate synthase [NADPH] large chain precursor (EC 1.4.1.13) (Glutamate synthase subunit alpha) (NADPH-GOGAT) (GLTS alpha chain) [Gene: gltB] - Azospirillum brasilense
1ofd	24.7	0.8	135	81	1.4.7.1	Ferredoxin-dependent glutamate synthase 2 (EC 1.4.7.1) (FD-GOGAT) [Gene: gltS or sll1499] - Synechocystis sp. (strain PCC 6803)
1rd5	5.7	2.6	77	13	4.2.1.20	Indole-3-glycerol phosphate lyase, chloroplast precursor (EC 4.1.2.8) (Indole synthase) (Tryptophan synthase alpha chain) (EC 4.2.1.20) (Benzoxazineless 1) [Gene: BX1] - Zea mays (Maize)
1k7f	5.2	2.3	63	14	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1q6o	5.2	2.6	75	15	4.1.1.85	3-keto-L-gulonate-6-phosphate decarboxylase ulaD (EC 4.1.1.85) (3-dehydro-L-gulonate-6-phosphate decarboxylase) (KGPDC) (L-ascorbate utilization protein D) [Gene: ulaD or sgaH or yjfV or b4196 or JW4154] - Escherichia coli (strain K12)
1qoq	5.1	2.3	63	14	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1k8y	5	2.4	65	14	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1k8z	4.9	2.5	64	14	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1k3u	4.9	2.7	71	15	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1ttp	4.8	2.3	62	15	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1k7x	4.8	2.4	63	14	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1ttq	4.7	2.2	61	15	4.2.1.20	Outer membrane protein tolC precursor [Gene: tolC or mtcB or mukA or refI or b3035 or JW5503] - Escherichia coli (strain K12)
1kfj	4.6	2.4	61	15	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1g7v	4.6	2.9	75	7	2.5.1.55	2-dehydro-3-deoxyphosphooctonate aldolase (EC 2.5.1.55) (Phospho-2-dehydro-3-deoxyoctonate aldolase) (3-deoxy-D-manno-octulosonic acid 8-phosphate synthetase) (KDO-8-phosphate synthetase) (KDO 8-P synthase) (KDOPS) [Gene: kdsA or b1215 or JW1206] - Escherichia coli (strain K12)
1ec7	4.5	2.3	62	15	4.2.1.40	Glucarate dehydratase (EC 4.2.1.40) (GDH) (GlucD) [Gene: gudD or ygcX or b2787 or JW2758] - Escherichia coli (strain K12)
1kfk	4.5	2.4	60	13	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1xc4	4.5	2.7	66	12	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or b1260 or JW1252] - Escherichia coli (strain K12)
1h7w	4.3	2.7	78	14	1.3.1.2	Dihydropyrimidine dehydrogenase [NADP+] precursor (EC 1.3.1.2) (DPD) (DHPDHase) (Dihydrouracil dehydrogenase) (Dihydrothymine dehydrogenase) [Gene: DPYD] - Sus scrofa (Pig)
1qw8	4.3	2.7	82	15	3.2.1.55	Alpha-N-arabinofuranosidase (EC 3.2.1.55) (Arabinosidase) [Gene: abfA] - Bacillus stearothermophilus (Geobacillus stearothermophilus)
1n82	4.2	3.1	88	9	not found	The high-resolution crystal structure of IXT6, a thermophilic, intracellular xylanase from G. stearothermophilus
1qw9	4.1	2.6	79	11	3.2.1.55	Alpha-N-arabinofuranosidase (EC 3.2.1.55) (Arabinosidase) [Gene: abfA] - Bacillus stearothermophilus (Geobacillus stearothermophilus)
1esw	28.4	0			2.4.1.25	4-alpha-glucanotransferase (EC 2.4.1.25) (Amylomaltase) (Disproportionating enzyme) (D-enzyme) [Gene: malQ] - Thermus thermophilus
1fp9	20.7	0.9	122	97	2.4.1.25	4-alpha-glucanotransferase (EC 2.4.1.25) (Amylomaltase) (Disproportionating enzyme) (D-enzyme) [Gene: malQ] - Thermus thermophilus
1fp8	19.5	0.9	113	96	2.4.1.25	4-alpha-glucanotransferase (EC 2.4.1.25) (Amylomaltase) (Disproportionating enzyme) (D-enzyme) [Gene: malQ] - Thermus thermophilus
1tz7	8.5	2.5	93	61	2.4.1.25	4-alpha-glucanotransferase (EC 2.4.1.25) (Amylomaltase) (Disproportionating enzyme) (D-enzyme) [Gene: malQ or malM or aq_723] - Aquifex aeolicus
1jg9	6.9	2.7	92	11	2.4.1.4	Amylosucrase (EC 2.4.1.4) [Gene: ams] - Neisseria polysaccharea
1g5a	6.9	3.2	97	11	2.4.1.4	Amylosucrase (EC 2.4.1.4) [Gene: ams] - Neisseria polysaccharea

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1jgi	6.8	3.3	93	10	2.4.1.4	Amylosucrase (EC 2.4.1.4) [Gene: ams] - Neisseria polysaccharea
1mw2	6.5	2.6	90	11	2.4.1.4	Amylosucrase (EC 2.4.1.4) [Gene: ams] - Neisseria polysaccharea
1mvy	6.4	2.6	92	10	2.4.1.4	Amylosucrase (EC 2.4.1.4) [Gene: ams] - Neisseria polysaccharea
1mw1	6.4	2.6	90	11	2.4.1.4	Amylosucrase (EC 2.4.1.4) [Gene: ams] - Neisseria polysaccharea
1m53	6.1	3	93	14	not found	CRYSTAL STRUCTURE OF ISOMALTULOSE SYNTHASE (PALI) FROM KLEBSIELLA SP. LX3
1izj	6	2.8	90	8	3.2.1.135	Neopullulanase 1 precursor (EC 3.2.1.135) (Alpha-amylase I) (TVA I) [Gene: tvaI] - Thermoactinomyces vulgaris
1izk	6	2.8	90	8	3.2.1.135	Neopullulanase 1 precursor (EC 3.2.1.135) (Alpha-amylase I) (TVA I) [Gene: tvaI] - Thermoactinomyces vulgaris
1ji1	5.6	2.7	86	8	3.2.1.135	Neopullulanase 1 precursor (EC 3.2.1.135) (Alpha-amylase I) (TVA I) [Gene: tvaI] - Thermoactinomyces vulgaris
1kfe	5.2	3	87	7	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or STM1727] - Salmonella typhimurium
1geq	4.8	2.8	90	8	4.2.1.20	Tryptophan synthase alpha chain (EC 4.2.1.20) [Gene: trpA or PF1705] - Pyrococcus furiosus
1etu	17.2	0			0.0.0.0	Elongation factor Tu (EF-Tu) (P-43) [Gene: (tufA or b3339 or JW3301) and (tufB or b3980 or JW3943)] - Escherichia coli (strain K12)
2bvn	10.8	1.3	68	94	0.0.0.0	Elongation factor Tu (EF-Tu) (P-43) [Gene: (tufA or b3339 or JW3301) and (tufB or b3980 or JW3943)] - Escherichia coli (strain K12)
1mj1	9.7	1.7	70	70	not found	30S ribosomal protein S12 [Gene: rpsL or strA or b3342 or JW3304] - Escherichia coli (strain K12)
1ob2	8.5	1.7	70	81	0.0.0.0	Elongation factor Tu (EF-Tu) (P-43) [Gene: (tufA or b3339 or JW3301) and (tufB or b3980 or JW3943)] - Escherichia coli (strain K12)
1fnm	8.5	2	65	34	0.0.0.0	Elongation factor G (EF-G) [Gene: fusA or fus] - Thermus thermophilus
1s1h	8.1	1.8	65	40	0.0.0.0	Elongation factor 2 (EF-2) (Translation elongation factor 2) (Eukaryotic elongation factor 2) (eEF2) (Ribosomal translocase) [Gene: (EFT1 or YOR133W or O3317 or YOR3317W) and (EFT2 or YDR385W)] - Saccharomyces cerevisiae (Baker's yeast)
1zm9	7.9	1.6	66	35	0.0.0.0	Elongation factor 2 (EF-2) (Translation elongation factor 2) (Eukaryotic elongation factor 2) (eEF2) (Ribosomal translocase) [Gene: (EFT1 or YOR133W or O3317 or YOR3317W) and (EFT2 or YDR385W)] - Saccharomyces cerevisiae (Baker's yeast)
1zm3	7.9	1.6	66	35	0.0.0.0	Elongation factor 2 (EF-2) (Translation elongation factor 2) (Eukaryotic elongation factor 2) (eEF2) (Ribosomal translocase) [Gene: (EFT1 or YOR133W or O3317 or YOR3317W) and (EFT2 or YDR385W)] - Saccharomyces cerevisiae (Baker's yeast)
1ktv	7.9	1.8	66	35	0.0.0.0	Elongation factor G (EF-G) [Gene: fusA or fus] - Thermus thermophilus
1n0v	7.6	1.8	65	32	0.0.0.0	Elongation factor 2 (EF-2) (Translation elongation factor 2) (Eukaryotic elongation factor 2) (eEF2) (Ribosomal translocase) [Gene: (EFT1 or YOR133W or O3317 or YOR3317W) and (EFT2 or YDR385W)] - Saccharomyces cerevisiae (Baker's yeast)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1s0u	7.5	1.2	62	53	0.0.0.0	Translation initiation factor 2 subunit gamma (eIF-2-gamma) (aIF2-gamma) [Gene: eif2g or MJ1261] - Methanocaldococcus jannaschii (Methanococcus jannaschii)
1g7s	7.5	2.3	60	43	0.0.0.0	Probable translation initiation factor IF-2 [Gene: infB or MTH_259] - Methanobacterium thermoautotrophicum
1n0u	7.4	1.8	65	40	0.0.0.0	Elongation factor 2 (EF-2) (Translation elongation factor 2) (Eukaryotic elongation factor 2) (eEF2) (Ribosomal translocase) [Gene: (EFT1 or YOR133W or O3317 or YOR3317W) and (EFT2 or YDR385W)] - Saccharomyces cerevisiae (Baker's yeast)
1wb1	6.5	1	51	51	not found	CRYSTAL STRUCTURE OF TRANSLATION ELONGATION FACTOR SELB FROM METHANOCOCCUS MARIPALUDIS IN COMPLEX WITH GDP
1tui	6.2	0.6	52	90	0.0.0.0	Elongation factor Tu (EF-Tu) [Gene: tuf or tufA] - Thermus aquaticus
1ije	5.8	2.8	57	51	0.0.0.0	Elongation factor 1-alpha (EF-1-alpha) (Translation elongation factor 1A) (Eukaryotic elongation factor 1A) (eEF1A) [Gene: (TEF1 or YPR080W or P9513.7) and (TEF2 or YBR118W or YBR0913)] - Saccharomyces cerevisiae (Baker's yeast)
1zun	5.2	1.9	54	50	not found	Sulfate adenylyltransferase subunit 2 (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase small subunit) [Gene: cysD or PSPTO_4433] - Pseudomonas syringae pv. tomato
1r5o	5	1.8	54	43	0.0.0.0	Eukaryotic peptide chain release factor GTP-binding subunit (ERF2) (Translation release factor 3) (Polypeptide release factor 3) (ERF3) (ERF-3) [Gene: sup35 or SPCC584.04] - Schizosaccharomyces pombe (Fission yeast)
1r5n	4.8	1.6	53	47	0.0.0.0	Eukaryotic peptide chain release factor GTP-binding subunit (ERF2) (Translation release factor 3) (Polypeptide release factor 3) (ERF3) (ERF-3) [Gene: sup35 or SPCC584.04] - Schizosaccharomyces pombe (Fission yeast)
1r5b	4.6	1.6	53	47	0.0.0.0	Eukaryotic peptide chain release factor GTP-binding subunit (ERF2) (Translation release factor 3) (Polypeptide release factor 3) (ERF3) (ERF-3) [Gene: sup35 or SPCC584.04] - Schizosaccharomyces pombe (Fission yeast)
1f60	4.6	1.8	49	49	0.0.0.0	Elongation factor 1-alpha (EF-1-alpha) (Translation elongation factor 1A) (Eukaryotic elongation factor 1A) (eEF1A) [Gene: (TEF1 or YPR080W or P9513.7) and (TEF2 or YBR118W or YBR0913)] - Saccharomyces cerevisiae (Baker's yeast)
1yzu	4.3	1.9	48	29	0.0.0.0	Ras-related protein Rab-21 [Gene: RAB21 or KIAA0118] - Homo sapiens (Human)
1euq	27.1	0			6.1.1.18	Glutaminyl-tRNA synthetase (EC 6.1.1.18) (Glutamine--tRNA ligase) (GlnRS) [Gene: glnS or b0680 or JW0666] - Escherichia coli (strain K12)
1euy	25	0.3	128	100	6.1.1.18	Glutaminyl-tRNA synthetase (EC 6.1.1.18) (Glutamine--tRNA ligase) (GlnRS) [Gene: glnS or b0680 or JW0666] - Escherichia coli (strain K12)
1gtr	23.7	0.4	126	100	6.1.1.18	Glutaminyl-tRNA synthetase (EC 6.1.1.18) (Glutamine--tRNA ligase) (GlnRS) [Gene: glnS or b0680 or JW0666] - Escherichia coli (strain K12)
1zjw	22.8	0.3	127	100	6.1.1.18	Glutaminyl-tRNA synthetase (EC 6.1.1.18) (Glutamine--tRNA ligase) (GlnRS) [Gene: glnS or b0680 or JW0666] - Escherichia coli (strain K12)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1qrt	21.7	0.4	125	100	6.1.1.18	Glutaminyl-tRNA synthetase (EC 6.1.1.18) (Glutamine--tRNA ligase) (GlnRS) [Gene: glnS or b0680 or JW0666] - Escherichia coli (strain K12)
1o0c	20.7	0.3	119	100	6.1.1.18	Glutaminyl-tRNA synthetase (EC 6.1.1.18) (Glutamine--tRNA ligase) (GlnRS) [Gene: glnS or b0680 or JW0666] - Escherichia coli (strain K12)
1nyl	16.1	1.4	116	99	6.1.1.18	Glutaminyl-tRNA synthetase (EC 6.1.1.18) (Glutamine--tRNA ligase) (GlnRS) [Gene: glnS or b0680 or JW0666] - Escherichia coli (strain K12)
1v47	9.7	2.7	116	16	not found	Crystal structure of ATP sulfurylase from Thermus thermophillus HB8 in complex with APS
1xjq	8.6	2.8	114	12	2.7.1.25	Bifunctional 3'-phosphoadenosine 5'-phosphosulfate synthetase 1 (PAPS synthetase 1) (PAPSS 1) (Sulfurylase kinase 1) (SK1) (SK 1) [Includes: Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase); Adenylyl-sulfate kinase (EC 2.7.1.25) (Adenylylsulfate 3'-phosphotransferase) (APS kinase) (Adenosine-5'-phosphosulfate 3'-phosphotransferase) (3'-phosphoadenosine-5'-phosphosulfate synthetase)] [Gene: PAPSS1 or ATPSK1 or PAPSS] - Homo sapiens (Human)
1xnj	8.4	2.9	114	11	2.7.1.25	Bifunctional 3'-phosphoadenosine 5'-phosphosulfate synthetase 1 (PAPS synthetase 1) (PAPSS 1) (Sulfurylase kinase 1) (SK1) (SK 1) [Includes: Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase); Adenylyl-sulfate kinase (EC 2.7.1.25) (Adenylylsulfate 3'-phosphotransferase) (APS kinase) (Adenosine-5'-phosphosulfate 3'-phosphotransferase) (3'-phosphoadenosine-5'-phosphosulfate synthetase)] [Gene: PAPSS1 or ATPSK1 or PAPSS] - Homo sapiens (Human)
1g8g	8.1	3.1	111	13	2.7.7.4	Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase) (Methionine-requiring protein 3) [Gene: MET3 or YJR010W or J1436] - Saccharomyces cerevisiae (Baker's yeast)
1jed	8	2.9	110	13	2.7.7.4	Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase) (Methionine-requiring protein 3) [Gene: MET3 or YJR010W or J1436] - Saccharomyces cerevisiae (Baker's yeast)
1g8h	8	2.9	110	13	2.7.7.4	Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase) (Methionine-requiring protein 3) [Gene: MET3 or YJR010W or J1436] - Saccharomyces cerevisiae (Baker's yeast)
1jhd	8	3	108	10	2.7.7.4	Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase) [Gene: sat or sopT] - Riftia pachyptila sulfur-oxidizing endosymbiont
1jee	7.9	2.9	110	13	2.7.7.4	Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase) (Methionine-requiring protein 3) [Gene: MET3 or YJR010W or J1436] - Saccharomyces cerevisiae (Baker's yeast)
1j70	7.9	2.9	112	13	2.7.7.4	Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase) (Methionine-requiring protein 3) [Gene: MET3 or YJR010W or J1436] - Saccharomyces cerevisiae (Baker's yeast)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1x6v	7.7	2.9	111	12	2.7.1.25	Bifunctional 3'-phosphoadenosine 5'-phosphosulfate synthetase 1 (PAPS synthetase 1) (PAPSS 1) (Sulfurylase kinase 1) (SK1) (SK 1) [Includes: Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase); Adenylyl-sulfate kinase (EC 2.7.1.25) (Adenylylsulfate 3'-phosphotransferase) (APS kinase) (Adenosine-5'-phosphosulfate 3'-phosphotransferase) (3'-phosphoadenosine-5'-phosphosulfate synthetase)] [Gene: PAPSS1 or ATPSK1 or PAPSS] - Homo sapiens (Human)
1i2d	7.7	3	109	16	2.7.7.4	Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase) [Gene: MET3 or APS] - Penicillium chrysogenum (Penicillium notatum)
1g8f	7.6	3.3	113	14	2.7.7.4	Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase) (Methionine-requiring protein 3) [Gene: MET3 or YJR010W or J1436] - Saccharomyces cerevisiae (Baker's yeast)
1r6x	7.4	3.1	109	13	2.7.7.4	Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase) (Methionine-requiring protein 3) [Gene: MET3 or YJR010W or J1436] - Saccharomyces cerevisiae (Baker's yeast)
1m8p	6.9	3.2	105	12	2.7.7.4	Sulfate adenylyltransferase (EC 2.7.7.4) (Sulfate adenylate transferase) (SAT) (ATP-sulfurylase) [Gene: MET3 or APS] - Penicillium chrysogenum (Penicillium notatum)
1fp1	26.4	0			2.1.1.-	Isoliquiritigenin 2'-O-methyltransferase (EC 2.1.1.-) (Chalcone O-methyltransferase) (ChOMT) - Medicago sativa (Alfalfa)
1kyz	18.3	1	110	49	2.1.1.68	Caffeic acid 3-O-methyltransferase (EC 2.1.1.68) (S-adenosysl-L-methionine:caffeic acid 3-O-methyltransferase) (COMT) (CAOMT) - Medicago sativa (Alfalfa)
1fpq	17.3	1.9	103	95	2.1.1.-	Isoliquiritigenin 2'-O-methyltransferase (EC 2.1.1.-) (Chalcone O-methyltransferase) (ChOMT) - Medicago sativa (Alfalfa)
1fp2	14.2	2.1	111	30	2.1.1.150	Isoflavone-7-O-methyltransferase 8 (EC 2.1.1.150) (Isoflavone-O-methyltransferase 8) (7-IOMT-8) - Medicago sativa (Alfalfa)
1tw3	13.6	2	109	27	2.1.1.-	Carminomycin 4-O-methyltransferase (EC 2.1.1.-) (COMT) [Gene: dnrK] - Streptomyces peucetius
1fpx	13.6	2.2	108	30	2.1.1.150	Isoflavone-7-O-methyltransferase 8 (EC 2.1.1.150) (Isoflavone-O-methyltransferase 8) (7-IOMT-8) - Medicago sativa (Alfalfa)
1r00	11.4	1.6	93	31	not found	Crystal structure of aclacinomycin-10-hydroxylase (RdmB) in complex with S-adensyl-L-homocystein (SAH)
2ex4	10	1.9	91	18	0.0.0.0	UPF0351 protein C9orf32 [Gene: C9orf32 or AD-003] - Homo sapiens (Human)
1im8	9.6	2.3	87	17	0.0.0.0	tRNA (cmo5U34)-methyltransferase (EC 2.1.1.-) [Gene: cmoA or HI0319] - Haemophilus influenzae
1vlm	9.4	2	95	16	not found	Crystal structure of SAM-dependent methyltransferase, possible histamine N-methyltransferase (TM1293) from Thermotoga maritima at 2.20 A resolution
1kph	8.5	2.1	84	18	2.1.1.79	Cyclopropane-fatty-acyl-phospholipid synthase 1 (EC 2.1.1.79) (Cyclopropane fatty acid synthase) (CFA synthase) (Cyclopropane mycolic acid synthase 1) [Gene: cmaA1 or cma1 or Rv3392c or MT3499 or MTV004.50] - Mycobacterium tuberculosis
1xxl	8.5	3.2	90	16	not found	The crystal structure of YcgJ protein from Bacillus subitilis at 2.1 A resolution

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1vl5	8.2	2.8	85	14	not found	Crystal structure of putative methyltransferase (BH2331) from Bacillus halodurans at 1.95 A resolution
1kpg	8.1	2.1	82	20	2.1.1.79	Cyclopropane-fatty-acyl-phospholipid synthase 1 (EC 2.1.1.79) (Cyclopropane fatty acid synthase) (CFA synthase) (Cyclopropane mycolic acid synthase 1) [Gene: cmaA1 or cma1 or Rv3392c or MT3499 or MTV004.50] - Mycobacterium tuberculosis
1xtp	7.9	2.2	83	18	not found	Structural Analysis of Leishmania major LMAJ004091AAA, a SAM-dependent methyltransferase of the DUF858/Pfam05891 family
2fk7	7.9	2.5	85	14	not found	Crystal structure of Hma (MmaA4) from Mycobacterium tuberculosis, apo-form
1kpi	7.5	2.5	93	20	2.1.1.79	Cyclopropane-fatty-acyl-phospholipid synthase 2 (EC 2.1.1.79) (Cyclopropane fatty acid synthase) (CFA synthase) (Cyclopropane mycolic acid synthase 2) [Gene: cmaA2 or cma2 or Rv0503c or MT0524 or MTCY20G9.30c] - Mycobacterium tuberculosis
1kia	6.7	3.7	90	14	2.1.1.20	Glycine N-methyltransferase (EC 2.1.1.20) (Folate-binding protein) [Gene: Gnmt] - Rattus norvegicus (Rat)
2avn	6.1	2.6	82	11	not found	Crystal structure of Ubiquinone/menaquinone biosynthesis methyltransferase-related protein (tm1389) from THERMOTOGA MARITIMA at 2.35 A resolution
1kp9	6	2.3	74	14	2.1.1.79	Cyclopropane-fatty-acyl-phospholipid synthase 1 (EC 2.1.1.79) (Cyclopropane fatty acid synthase) (CFA synthase) (Cyclopropane mycolic acid synthase 1) [Gene: cmaA1 or cma1 or Rv3392c or MT3499 or MTV004.50] - Mycobacterium tuberculosis
1ic3	5.7	2.8	75	13	not found	mRNA cap guanine-N7 methyltransferase (EC 2.1.1.56) (mRNA (guanine-N(7)-)-methyltransferase) (mRNA cap methyltransferase) [Gene: ABD1 or YBR236C or YBR1602] - Saccharomyces cerevisiae (Baker's yeast)
1ri2	5.3	2.8	84	14	2.1.1.56	mRNA cap guanine-N7 methyltransferase (EC 2.1.1.56) (mRNA (guanine-N(7)-)-methyltransferase) (mRNA cap methyltransferase) [Gene: ABD1 or ECU10_0380] - Encephalitozoon cuniculi
1r8y	4.3	2.2	66	17	2.1.1.20	Glycine N-methyltransferase (EC 2.1.1.20) [Gene: Gnmt] - Mus musculus (Mouse)
1in4	25.4	0			3.6.1.-	Holliday junction ATP-dependent DNA helicase ruvB (EC 3.6.1.-) [Gene: ruvB or TM_1730] - Thermotoga maritima
1njf	10.9	1.5	87	31	2.7.7.7	DNA polymerase III subunit tau (EC 2.7.7.7) [Contains: DNA polymerase III subunit gamma] [Gene: dnaX or dnaZ or dnaZX or b0470 or JW0459] - Escherichia coli (strain K12)
1njf	10.8	1.5	87	31	2.7.7.7	DNA polymerase III subunit tau (EC 2.7.7.7) [Contains: DNA polymerase III subunit gamma] [Gene: dnaX or dnaZ or dnaZX or b0470 or JW0459] - Escherichia coli (strain K12)
1sxj	10.2	2.1	93	26	0.0.0.0	Proliferating cell nuclear antigen (PCNA) [Gene: POL30 or YBR088C or YBR0811] - Saccharomyces cerevisiae (Baker's yeast)
1njg	9.5	1.3	77	36	2.7.7.7	DNA polymerase III subunit tau (EC 2.7.7.7) [Contains: DNA polymerase III subunit gamma] [Gene: dnaX or dnaZ or dnaZX or b0470 or JW0459] - Escherichia coli (strain K12)
1sxj	7.6	2.2	76	26	0.0.0.0	Proliferating cell nuclear antigen (PCNA) [Gene: POL30 or YBR088C or YBR0811] - Saccharomyces cerevisiae (Baker's yeast)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1njf	4.8	1.1	61	41	2.7.7.7	DNA polymerase III subunit tau (EC 2.7.7.7) [Contains: DNA polymerase III subunit gamma] [Gene: dnaX or dnaZ or dnaZX or b0470 or JW0459] - Escherichia coli (strain K12)
1jr3	21.9	0			2.7.7.7	DNA polymerase III subunit tau (EC 2.7.7.7) [Contains: DNA polymerase III subunit gamma] [Gene: dnaX or dnaZ or dnaZX or b0470 or JW0459] - Escherichia coli (strain K12)
1njg	13.1	2.1	105	97	2.7.7.7	DNA polymerase III subunit tau (EC 2.7.7.7) [Contains: DNA polymerase III subunit gamma] [Gene: dnaX or dnaZ or dnaZX or b0470 or JW0459] - Escherichia coli (strain K12)
1njf	11.8	1.3	88	100	2.7.7.7	DNA polymerase III subunit tau (EC 2.7.7.7) [Contains: DNA polymerase III subunit gamma] [Gene: dnaX or dnaZ or dnaZX or b0470 or JW0459] - Escherichia coli (strain K12)
1njf	11.6	1.7	88	99	2.7.7.7	DNA polymerase III subunit tau (EC 2.7.7.7) [Contains: DNA polymerase III subunit gamma] [Gene: dnaX or dnaZ or dnaZX or b0470 or JW0459] - Escherichia coli (strain K12)
1ka9	11.7	0			2.4.2.-	Imidazole glycerol phosphate synthase subunit hisH (EC 2.4.2.-) (IGP synthase glutamine amidotransferase subunit) (IGP synthase subunit hisH) (ImGP synthase subunit hisH) (IGPS subunit hisH) [Gene: hisH or TTHA0430] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1nal	4.4	1.4	37	24	4.1.3.3	N-acetylneuraminate lyase (EC 4.1.3.3) (N-acetylneuraminic acid aldolase) (N-acetylneuraminate pyruvate-lyase) (Sialic acid lyase) (Sialate lyase) (Sialic acid aldolase) (NALase) [Gene: nanA or npl or b3225 or JW3194] - Escherichia coli (strain K12)
1mzh	4.1	1.9	47	11	4.1.2.4	Deoxyribose-phosphate aldolase (EC 4.1.2.4) (Phosphodeoxyriboaldolase) (Deoxyriboaldolase) (DERA) [Gene: deoC or aq_148] - Aquifex aeolicus
2b7p	4.1	2.5	48	21	2.4.2.19	Probable nicotinate-nucleotide pyrophosphorylase [carboxylating] (EC 2.4.2.19) (Quinolinate phosphoribosyltransferase [decarboxylating]) (QAPRTase) [Gene: nadC or HP_1355] - Helicobacter pylori (Campylobacter pylori)
1llw	31.7	0			1.4.7.1	Ferredoxin-dependent glutamate synthase 2 (EC 1.4.7.1) (FD-GOGAT) [Gene: gltS or sll1499] - Synechocystis sp. (strain PCC 6803)
1lm1	29.8	0.2	139	100	1.4.7.1	Ferredoxin-dependent glutamate synthase 2 (EC 1.4.7.1) (FD-GOGAT) [Gene: gltS or sll1499] - Synechocystis sp. (strain PCC 6803)
1llz	29.1	0.3	138	100	1.4.7.1	Ferredoxin-dependent glutamate synthase 2 (EC 1.4.7.1) (FD-GOGAT) [Gene: gltS or sll1499] - Synechocystis sp. (strain PCC 6803)
1ofe	27.9	0.4	139	100	1.4.7.1	Ferredoxin-dependent glutamate synthase 2 (EC 1.4.7.1) (FD-GOGAT) [Gene: gltS or sll1499] - Synechocystis sp. (strain PCC 6803)
1xl9	4.8	2.3	68	13	not found	Crystal Structure of Dihydrodipicolinate Synthase DapA-2 (BA3935) from Bacillus Anthracis.
1ub3	4.5	2.2	68	16	not found	Deoxyribose-phosphate aldolase (EC 4.1.2.4) (Phosphodeoxyriboaldolase) (Deoxyriboaldolase) (DERA) [Gene: deoC or TTHA1186] - Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)
1lor	4.5	2.5	62	13	4.1.1.23	Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23) (OMP decarboxylase) (OMPDCase) (OMPdecase) [Gene: pyrF or MTH_129] - Methanobacterium thermoautotrophicum

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1xi3	4.4	1.8	56	13	2.5.1.3	Thiamine-phosphate pyrophosphorylase (EC 2.5.1.3) (TMP pyrophosphorylase) (TMP-PPase) (Thiamine-phosphate synthase) [Gene: thiE or PF1334] - Pyrococcus furiosus
1p1x	4.4	2.4	73	7	not found	Deoxyribose-phosphate aldolase (EC 4.1.2.4) (Phosphodeoxyriboaldolase) (Deoxyriboaldolase) (DERA) [Gene: deoC or dra or thyR or b4381 or JW4344] - Escherichia coli (strain K12)
1sja	4.3	2.1	66	21	not found	X-ray structure of o-Succinylbenzoate Synthase complexed with N-acetylmethionine
1sjb	4.3	2.1	66	21	not found	X-ray structure of o-succinylbenzoate synthase complexed with o-succinylbenzoic acid
1sjc	4.3	2.1	66	20	not found	x-ray structure of o-succinylbenzoate synthase complexed with N-succinyl methionine
1sjd	4.3	2.1	66	21	not found	x-ray structure of o-succinylbenzoate synthase complexed with n-succinyl phenylglycine
1yad	4.2	1.8	59	10	0.0.0.0	Regulatory protein tenI [Gene: tenI or BSU11660] - Bacillus subtilis
1tww	4.2	2.1	64	17	not found	Dihydropteroate Synthetase, With Bound Substrate Analogue PtPP, From Bacillus anthracis
1tww	4.2	2.1	64	17	not found	Dihydropteroate Synthetase, With Bound Substrate Analogue PtPP, From Bacillus anthracis
1vlw	4.2	2.3	63	13	not found	Crystal structure of 2-dehydro-3-deoxyphosphogluconate aldolase/4-hydroxy-2-oxoglutarate aldolase (TM0066) from Thermotoga maritima at 2.30 A resolution
1loq	4.2	2.5	63	14	4.1.1.23	Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23) (OMP decarboxylase) (OMPDCase) (OMPdecase) [Gene: pyrF or MTH_129] - Methanobacterium thermoautotrophicum
1o0y	4.1	2	66	12	4.1.2.4	Deoxyribose-phosphate aldolase (EC 4.1.2.4) (Phosphodeoxyriboaldolase) (Deoxyriboaldolase) (DERA) [Gene: deoC or TM_1559] - Thermotoga maritima
1q6l	4.1	2.8	72	8	4.1.1.85	3-keto-L-gulonate-6-phosphate decarboxylase ulaD (EC 4.1.1.85) (3-dehydro-L-gulonate-6-phosphate decarboxylase) (KGPDC) (L-ascorbate utilization protein D) [Gene: ulaD or sgaH or yjfV or b4196 or JW4154] - Escherichia coli (strain K12)
1pjq	18	0			2.1.1.107	Siroheme synthase [Includes: Uroporphyrinogen-III C-methyltransferase (EC 2.1.1.107) (Urogen III methylase) (SUMT) (Uroporphyrinogen III methylase) (UROM); Precorrin-2 dehydrogenase (EC 1.3.1.76); Sirohydrochlorin ferrochelatase (EC 4.99.1.4)] [Gene: cysG or STM3477] - Salmonella typhimurium
1pjs	16.9	0.2	72	100	2.1.1.107	Siroheme synthase [Includes: Uroporphyrinogen-III C-methyltransferase (EC 2.1.1.107) (Urogen III methylase) (SUMT) (Uroporphyrinogen III methylase) (UROM); Precorrin-2 dehydrogenase (EC 1.3.1.76); Sirohydrochlorin ferrochelatase (EC 4.99.1.4)] [Gene: cysG or STM3477] - Salmonella typhimurium
1s2l	5.1	2.2	59	10	not found	Purine 2'deoxyribosyltransferase native structure
1s2g	4.8	2.2	59	10	not found	Purine 2'deoxyribosyltransferase + 2'-deoxyadenosine
1s2i	4.8	2.2	59	10	not found	Purine 2'deoxyribosyltransferase + bromopurine
1s3f	4.8	2.2	59	10	not found	Purine 2'-deoxyribosyltransferase + selenoinosine
1s2d	4.8	2.3	58	10	not found	Purine 2'-Deoxyribosyl complex with arabinoside: Ribosylated Intermediate (AraA)

PDB Code Core	ZScore	RMSD	Alignment Length	Seqence Identity %	EC	Compound
1r7a	29.4	0			not found	Sucrose Phosphorylase from Bifidobacterium adolescentis
1tcm	4.4	2.1	74	19	2.4.1.19	Cyclomaltodextrin glucanotransferase precursor (EC 2.4.1.19) (Cyclodextrin-glycosyltransferase) (CGTase) [Gene: cgt] - Bacillus circulans
1v3l	4.2	1.8	70	20	2.4.1.19	Cyclomaltodextrin glucanotransferase precursor (EC 2.4.1.19) (Cyclodextrin-glycosyltransferase) (CGTase) [Gene: cgt] - Bacillus sp. (strain 1011)
1uks	4.1	1.8	70	21	2.4.1.19	Cyclomaltodextrin glucanotransferase precursor (EC 2.4.1.19) (Cyclodextrin-glycosyltransferase) (CGTase) [Gene: cgt] - Bacillus sp. (strain 1011)
1v3k	4.1	1.8	70	20	2.4.1.19	Cyclomaltodextrin glucanotransferase precursor (EC 2.4.1.19) (Cyclodextrin-glycosyltransferase) (CGTase) [Gene: cgt] - Bacillus sp. (strain 1011)
1su3	17.5	0			3.4.24.7	Interstitial collagenase precursor (EC 3.4.24.7) (Matrix metalloproteinase-1) (MMP-1) (Fibroblast collagenase) [Contains: 22 kDa interstitial collagenase; 27 kDa interstitial collagenase] [Gene: MMP1 or CLG] - Homo sapiens (Human)
1utt	9.8	0.5	70	63	3.4.24.65	Macrophage metalloelastase precursor (EC 3.4.24.65) (HME) (Matrix metalloproteinase-12) (MMP-12) (Macrophage elastase) (ME) [Gene: MMP12 or HME] - Homo sapiens (Human)
1you	7.7	0.4	69	58	3.4.24.-	Collagenase 3 precursor (EC 3.4.24.-) (Matrix metalloproteinase-13) (MMP-13) [Gene: MMP13] - Homo sapiens (Human)

Acknowledgements

I am very thankful to Prof. Dr. Herbert Waldmann for his supervision, for his encouragement, and for granting me the freedom to pursue my own ideas.

The move from synthetic chemistry to the field of cheminformatics was greatly facilitated by my mentors Dr. Peter Ertl, Prof. Dr. Tudor I. Oprea, and Dr. Ansgar Schuffenhauer. I am indebted to them for their continuous support and the many valuable lessons they have taught me. In this respect, I also express my thanks to my former colleague Dr. Steffen Renner for the numerous valuable discussions.

Dr. Daniel Rauh is also acknowledged for the many helpful discussions, and for his encouragement and support, especially within the Centre of Applied Chemical Genomics (ZACG) project.

The interdisciplinary projects described herein would hardly have been possible without contributions from a whole host of computer scientists, chemists and biologists. I am grateful to them all, especially to Drs. Peter Ertl, Ansgar Schuffenhauer and Silvio Roggo from the Novartis Institutes for Biomedical Research in Basel, to Dipl.-Inform. Karsten Klein and Prof. Dr. Petra Mutzel from the Faculty of Computer Science at the Technical University of Dortmund, to Dipl.-Chem. Anke Roth and Prof. Dr. Arenz from the Chemistry Department at the Humboldt University of Berlin, as well as to Bianca Sperl and Dr. Thorsten Berg at the Max Planck Institute of Biochemistry in Munich. I thank the members of the PG504 - Arbia Ben Ahmed, Anke Arndt, Philipp Büdenbender, Vanessa Bembenek, Adalbert Gorecki, Nils Kriege, Sergej Rakov, Michael Rex, Gebhard Schrader, Henning Wagner, André Wiesniewski and Cengizhan Yücel - for their enthusiasm and excellent collaboration in the development of Scaffold Hunter.

At the Max Planck Institute, I am greatly indebted to Dr. Christoph Schwittek for his continuous and excellent IT support, to Dr. Ingrid Vetter for sharing her expertise in protein structures, as well as to pharmacist Sabine Klueter and Dipl.-Biochem. Bernhard Ellinger for their introduction to the field of biochemical assays.

The Waldmann group is acknowledged for providing such a pleasant working atmosphere, for the constructive collaboration, helpful discussion, and for the fun had during the past years. In particular I wish to thank past and present office colleagues, as well as Drs. Lars Arve and Robin Bon, Dipl.-Biochem. Bernhard Ellinger, and Drs. Frank Dekker, and Thilo Walther. Special thanks must go to Dr. Christian Hedberg for the culinary heights and excellent wines.

Drs. Robin Bon, Lech-Gustav Milroy, and Thilo Walther are acknowledged for their quick but very thorough proof-reading of this manuscript and numerous helpful comments.

I am also very grateful to the numerous ‘invisible’ helpers behind the scenes who made this work possible but are too many to be named individually.

Last but certainly not least I thank my family and my friends for their understanding and support during the past years. This work is as much their success as it is mine.

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe.

Dortmund, 28.04.2009

_____

Stefan Wetzel